文法誤り訂正の包括的メタ評価: 既存自動評価の限界と大規模言語モデルの可能性

小林正宗 [1]　三田雅人 [2,1]　小町守 [3]
[1] 東京都立大学　[2] 株式会社サイバーエージェント　[3] 一橋大学
kobayashi-masamune@ed.tmu.ac.jp　mita_masato@cyberagent.co.jp
mamoru.komachi@r.hit-u.ac.jp

概要

評価尺度の評価（メタ評価）は主に人手評価との相関に基づいて行われる．しかし，従来の英語文法誤り訂正 (GEC) のメタ評価は，評価粒度の不一致によるバイアスや，現在の主流システムとの乖離などの問題に直面している．これらの問題に対処するために，12 の最先端システムを対象に，2 つの評価粒度で人手評価することで，より妥当性のあるメタ評価を可能にするデータセット (SEEDA) を構築する．さらに，SEEDA を用いた包括的なメタ評価を通して，GEC における既存自動評価の現状と大規模言語モデルの評価性能を調査する．

1　はじめに

文法誤り訂正 (GEC) の評価尺度は，その評価粒度によって編集ベースと文ベースの 2 つに分類され，それぞれ異なる目的を持っている．編集ベース評価尺度 (EBM) は編集自体を評価する一方，文ベース評価尺度 (SBM) は文全体の品質を評価する．また，そのような評価尺度の信頼性に関する評価（メタ評価）は，主に人手評価との相関に基づいて行われる．

英語 GEC のメタ評価の多くは，Grundkiewicz ら [1] のデータセット（GJG15）を使用してきたが，GJG15 に基づいたメタ評価 [1, 2, 3, 4] にはいくつかの問題がある．1 つ目に，評価尺度と人手評価の評価粒度の不一致から生じるバイアスにより，EBM が過小評価されている恐れがある．バイアスの例として，未訂正の文に対し，EBM は必ず最低のスコアをつける一方，文ベースの人手評価では文自体の良さを評価するため全範囲のスコアが付与され得ることが挙げられる．2 つ目に，GJG15 は CoNLL-2014 Shared Task [5] の統計的機械学習に基づく古典的システムを人手評価している．そのため，GJG15 のシステムと近年のディープニューラルネットワークに基づくシステムとの乖離により，メタ評価の適用範囲が制限されている．3 つ目に，単一のデータポイントから得られた相関に基づく結論は妥当性に欠ける可能性がある．Deutsch ら [6] は要約のメタ評価において特定の評価尺度は相関に幅があることを示しており，Mathur ら [7] は機械翻訳のメタ評価において外れ値が相関に強い影響を与えていたことを発見した．そのため，GEC においても同様の懸念が残る．

そこで本研究では，より信頼できる英語 GEC のメタ評価のための新たなデータセットである SEEDA を提案する．[1)] そして，構築した SEEDA を用いた包括的なメタ評価の結果，既存自動評価はこれまで報告されていたほど高性能システムの性能差を捉えるほどの解像度を持っていないことが明らかとなった．さらに，この現状を踏まえ，近年言語生成の評価モデルとしても有効と報告 [8, 9] されている大規模言語モデル (LLM) を GEC 評価にも初めて導入し，どの程度人手評価の代替になり得るか調査する．

2　関連研究

英語 GEC のメタ評価に関して，いくつかの研究が存在する．Grundkiewicz ら [1] と Napoles ら [10] は古典的システムに対して文ベースの人手評価を備えたデータセットを構築した．その後，Chollampatt ら [2] は評価尺度間の有意差検定の必要性を示し，Chosen ら [11] は人手評価に依存しないメタ評価方法を提案している．さらに，Napoles ら [12] は複数のドメインで人手評価を行うことで GMEG-Data を構築し，ドメイン間で多様な相関関係があることを示した．GMEG-Data は様々なドメインや CoNLL-2014 全体でのメタ評価ができる一方，SEEDA は近年の

1) Sentence-based and Edit-based human Evaluation DAtaset for GEC の略称．本データセットは今後一般公開する予定である．

表 1 編集ベースの人手評価の例．Step 1 では原文中の誤りは黄色，有効な編集は緑，無効な編集は赤で表示している．Step 2 では有効に訂正できた誤りは緑，訂正できなかった誤りは赤で表示している．

Step 1	**原文:** It is [againt] his or her human rights [] and it is against the [law's spirit]. **訂正文:** It is [againt → again] his or her human rights [→,] and it is against the [law's →] spirit [→ of the law].
Step 2	**原文:** It is [againt] his or her human rights [] and it is against the [law's spirit].
スコア	$F_{0.5}$ = **0.73**, Precision = 0.75, Recall = 0.67

高性能システムや評価粒度を考慮することで，より高い妥当性を提供している．

3 SEEDA データセット

3.1 文法誤り訂正システム

メタ評価における 2 つ目の問題点に対応するために，既存ベンチマークで高性能である 12 のシステムを考慮する [13]．具体的には，TemplateGEC [14]，TransGEC [15]，T5 [16]，LM-Critic [17]，BART [18]，BERT-fuse [19]，Riken Tohoku [20]，UEDIN-MS [21]，GECToR-ens [22]，GECToR-BERT [23]，PIE [24] および 2-shot の GPT-3.5 [25] を考慮する．また，GEC の評価では未訂正の文も考慮する必要があるため，CoNLL-2014 の原文 (INPUT) も含めた．さらに，システムの訂正性能を人手の訂正と比較するために，Sakaguchi ら [26] が作成した，英語母語話者かつ GEC の専門家による最小限の編集をしたリファレンス (REF-M) と流暢な文になるように編集をしたリファレンス (REF-F) も考慮し，計 15 の文集合を人手評価の対象とする．サンプリングは Grundkiewicz ら [1] が提案した多様な出力を好むように調整された分布を使用し，文集合から 200 のサブセットを収集した．ここで，人手評価の一致率を測定するために，サブセットの少なくとも 12.5 %を重複させ，1 つのサブセットには最大で 5 つの訂正文を含む．

3.2 アノテーションスキーム

SEEDA は編集ベースの人手評価 (SEEDA-E) と文ベースの人手評価 (SEEDA-S) の 2 種類から構成される．表 1 に示す通り，編集ベースの人手評価では，訂正文の編集に対して段階的な系列ラベリングを行う．Step 1 では，誤りカテゴリ [27] に従って原文の誤りを特定した後に，訂正文の各編集が有効かどうかを 2 値分類する．Step 2 では，各編集が Step 1 で見つかった誤りを訂正できたかを 2 値分類する．そして最後に，各訂正文の Precision と Recall から $F_{0.5}$[2] を算出する．また，文ベースの人手評価は GJG15 と同様の方法で行う．各アノテーションは，言語に関する仕事（翻訳，ネイティブチェック）に従事している 3 人の英語母語話者によって行われた．

2) GEC の自動評価では $F_{0.5}$ が一般的に使用される．

表 2 ペアワイズ評価の統計量．展開済みは重複した訂正文を展開した場合の値である．括弧内の値は同順位の数であり，左側が編集ベース，右側が文ベースを指す．

アノテータ	未展開	展開済み
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)
合計	5,347 (1,457 / 791)	33,544 (18,974 / 15,797)

表 3 未展開のペアワイズ評価を用いたアノテータ間およびアノテータ内のカッパ係数 (κ)．

	κ (SEEDA/GJG15)	程度
アノテータ間（編集）	0.28 / -	Fair
アノテータ間（文）	0.41 / 0.29	Moderate
アノテータ内（編集）	0.61 / -	Substantial
アノテータ内（文）	0.71 / 0.46	Substantial

3.3 分析

表 2 は，アノテータによる評価結果を元に作成したペアワイズ評価の統計量である．ペアワイズ評価は，2 つの文 (A, B) の全ての組み合わせをランキング ($A > B$, $A = B$, $A < B$) に取ることで作成した．展開済みのペアワイズ評価の数が未展開の数より大幅に多いことは，Grunkiewicz ら [1] と同様に GEC システムの出力には重複が多いことを示唆している．また，表 3 はアノテータ一致率であり，一致率としてカッパ係数 (κ) を使用した．本データセットの全ての一致率は，GJG15 と比較して高いことがわかった．さらに，表 5（付録 A）にペアワイズ評価から作成した人手ランキングを示す．GPT および T5 アーキテクチャに基づくシステム（GPT-3.5，T5，TransGEC）が REF-M よりも高い順位を獲得したことは，これらのシステムが人間より有効な訂正ができる可能性を示唆している．加えて，評価粒度間でのアノテータ内の平均 κ を算出したところ，文レベルでは評価粒度による人手評価の差が顕著 (κ =

表 4 外れ値を除いたシステムレベルおよび文レベルのメタ評価結果．太字のスコアは最も相関の高い評価尺度である．

評価尺度	システムレベル						文レベル					
	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**
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
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
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
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
LLM-E	-	-	0.839	0.846	0.911	0.965	-	-	0.698	0.395	0.728	0.455
LLM-E + Difficulty	-	-	**0.885**	**0.860**	**0.941**	0.972	-	-	0.717	0.434	0.719	0.437
LLM-E + Impact	-	-	0.844	**0.860**	0.905	**0.986**	-	-	**0.717**	**0.434**	**0.730**	**0.460**
LLM-S	-	-	0.887	0.860	0.960	0.958	-	-	0.784	0.567	0.798	0.595
LLM-S + Grammaticality	-	-	0.888	0.867	0.961	0.937	-	-	0.796	0.592	0.807	0.615
LLM-S + Fluency	-	-	**0.913**	0.874	**0.974**	**0.979**	-	-	**0.819**	**0.637**	**0.831**	**0.662**
LLM-S + Meaning Preservation	-	-	0.911	**0.881**	0.960	0.958	-	-	0.810	0.620	0.813	0.626

0.36) である一方，システムレベルでは比較的小さかった ($\kappa = 0.83$)．そのため，従来の文レベルのメタ評価では評価粒度の不一致によるバイアスがより顕著であった可能性が高い．

4 メタ評価

4.1 実験設定

評価粒度およびシステム集合のバイアスが緩和された新たなメタ評価用データセット SEEDA を用いた包括的メタ評価を通して，既存自動評価の現状および GEC 評価における LLM の可能性を明らかにする．調査対象の評価尺度として，EBM は M^2 [28]，ERRANT [27]，PT-M^2 [4]，GoToScorer [29] を，SBM は GLEU [10]，Scribendi Score [30]，SOME [3]，IMPARA [31] を考慮する．LLM は GPT-4 (`gpt-4-1106-preview`) を使用し，プロンプトによる評価性能の変化を調査するために，GEC の評価観点ごとのプロンプトを作成し，ベースプロンプトと比較する．編集ベースで評価する LLM (LLM-E) は訂正の難易度 (Difficulty) [29] と編集の影響度 (Impact) [31] に焦点を当て，文ベースで評価する LLM (LLM-S) は文法性 (Grammaticality) [32, 3] と流暢性 (Fluency) [32, 3]，意味保存性 (Meaning Preservation) [32, 3] に着目したプロンプトを使用する．各プロンプトの詳細は付録 B に記載する．システムレベルのメタ評価ではピアソンの相関（r）とスピアマンの順位相関（ρ），文レベルのメタ評価では Accuracy（Acc）とケンドールの順位相関（τ）を使用する．また，評価データを揃えるために、人手評価されたサブセットを用いて評価する．なお，INPUT と流暢さを目的とした訂正文（REF-F，GPT-3.5）は，外れ値として機能していたため，それらを除外した 12 システムを用いる．

4.2 既存自動評価の現状

評価粒度による影響: SEEDA-S と SEEDA-E を比較すると，文レベルのメタ評価において評価尺度と人手評価の評価粒度を一致させることで，相関が向上する傾向にあった (表 4)．つまり，評価粒度が一致していない場合，文レベルのメタ評価では評価尺度が過小評価されることになる．この要因として，文レベルでは評価粒度ごとに人手評価の結果が顕著に異なっていたことが挙げられる (§3.3)．

対象システムの影響: GJG15 と SEEDA-S を比較すると，対象システムが古典的システムから最先端ニューラルシステムに変わることで，相関が減少する傾向にあることがわかる．そのため，既存の評価尺度のほとんどは，高性能なニューラルシステムの訂正文をこれまでに報告されていたほど適切には評価できないことが確認された．

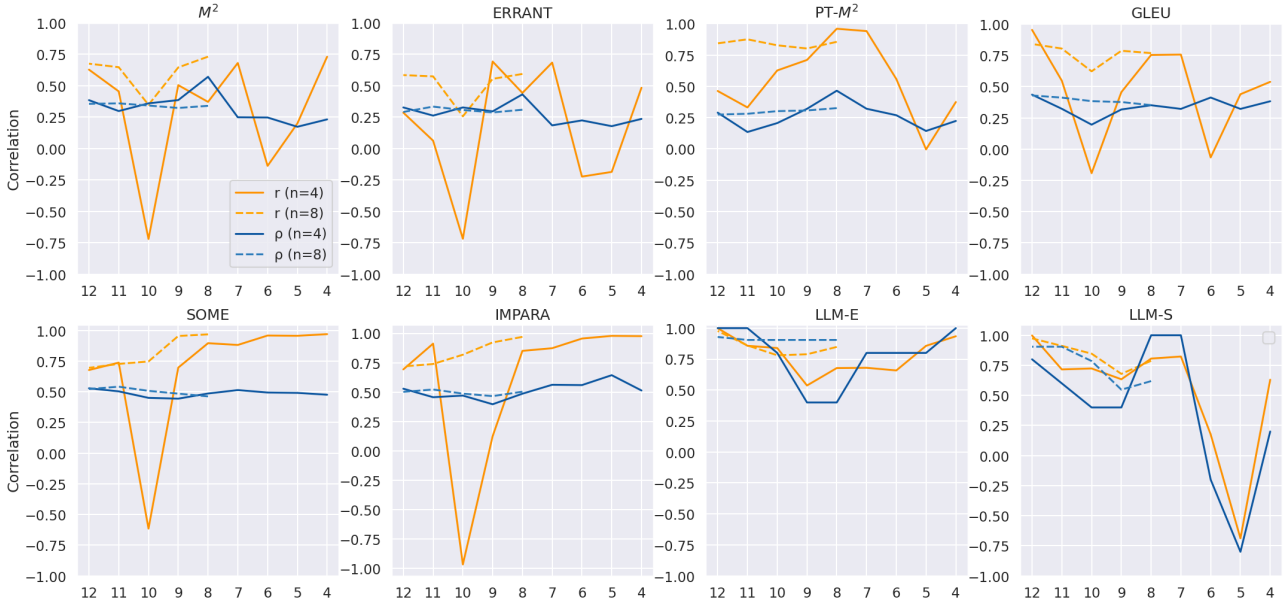

図 1 Window Analysis を用いてシステム集合を変化させた場合の相関の変動．x 軸は外れ値を除いた 12 のシステムの人手ランキングである．“n” は考慮されるシステムの数を示し，実線は 4 つのシステムを，破線は 8 つのシステムを表している．たとえば，$n = 4$ の場合，$x = 5$ の軸は人手ランキングで 2 位から 5 位のシステム集合に対応する．

複数の設定での相関の変動: 様々なシステム集合に対する相関を測るために Window Analysis（図 1）を行った．多くの評価尺度に共通していることは，相関はシステムが 4 つの場合は大きく変動する一方，8 つの場合は比較的に安定していることである．これは，既存の評価尺度が高性能システム間の性能差を捉えるのに十分な解像度を持っていないことを示唆している．さらに，M^2 や ERRANT，GLEU は頻繁に無相関か負の相関になることから，近年の GEC 評価は妥当性に欠けていた可能性がある．

4.3 LLM 評価モデルの可能性

メタ評価の結果，LLM は既存自動評価と比較して相関が高い傾向にあり，GEC の評価においても有用だとわかった (表 4)．観点ごとのプロンプトについては，ベースプロンプトと比較して相関を向上させる傾向にあった．特に，LLM-S + Fluency では文レベルの相関が大きく向上し，LLM の性能を大きく引き出すことができた．このことは，高性能システムを評価する際は文法性を超えて流暢性を精緻に見る必要があることを示唆している．また，システムレベルの LLM の相関の多くは 0.9 を超え，相関の比較が困難なほど高いため，システムレベルのメタ評価のような十数システム使って相関を算出するタスクは性能が飽和しつつあると考えられる。したがって，今後は文レベルの相関に着目したり，システムの数や組み合わせを柔軟に変えることで，タスクの難易度を上げることが必要である。

Window Analysis によると，LLM は比較的に高い相関を維持し相関の変動も小さいことから，近年の最先端システムを評価するのにより適していると考えられる．特に，EBM や SBM ではうまく評価できなかった 7 位から 10 位 (x=10) のシステム群に対して相関が高いことは，既存自動評価では評価が難しい領域を LLM が評価できる可能性を示唆している．

5 おわりに

本研究では，評価粒度およびシステム集合におけるバイアスが緩和されたより妥当性のあるメタ評価用データセット SEEDA を構築した．データセット分析の結果，評価粒度ごとに文レベルの人手評価結果は大きく異なり，GPT と T5 に基づく GEC システムは人間と同等以上の訂正ができることが明らかになった．相関分析の結果からは，従来のメタ評価手法では EBM の過小評価の可能性があること，文レベルのメタ評価では評価粒度を揃えると相関が改善される傾向にあることを示した．また，多くの既存自動評価は高性能システムの巧拙を捉えるほどの解像度を持っていないことが明らかとなった．さらに，LLM 評価モデルは既存自動評価と比較して相関が高く，GEC 評価での有用性が示された．

参考文献

[1] Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. Human evaluation of grammatical error correction systems. In **ACL**, 2015.

[2] Shamil Chollampatt and Hwee Tou Ng. A reassessment of reference-based grammatical error correction metrics. In **COLING**, 2018.

[3] Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. SOME: Reference-less submetrics optimized for manual evaluations of grammatical error correction. In **COLING**, 2020.

[4] Peiyuan Gong, Xuebo Liu, Heyan Huang, and Min Zhang. Revisiting grammatical error correction evaluation and beyond. In **EMNLP**, 2022.

[5] Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 shared task on grammatical error correction. In **CoNLL**, 2014.

[6] Daniel Deutsch, Rotem Dror, and Dan Roth. A statistical analysis of summarization evaluation metrics using resampling methods. **TACL**, 2021.

[7] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In **ACL**, 2020.

[8] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In **ACL**, 2023.

[9] Yixin Liu, Alexander R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization, 2023.

[10] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. Ground truth for grammatical error correction metrics. In **IJCNLP**, 2015.

[11] Leshem Choshen and Omri Abend. Automatic metric validation for grammatical error correction. In **ACL**, 2018.

[12] Courtney Napoles, Maria Nădejde, and Joel Tetreault. Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses. **TACL**, 2019.

[13] Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. Grammatical error correction: A survey of the state of the art. **CL**, 2023.

[14] Yinghao Li, Xuebo Liu, Shuo Wang, Peiyuan Gong, Derek F. Wong, Yang Gao, Heyan Huang, and Min Zhang. TemplateGEC: Improving grammatical error correction with detection template. In **ACL**, 2023.

[15] Tao Fang, Xuebo Liu, Derek F. Wong, Runzhe Zhan, Liang Ding, Lidia S. Chao, Dacheng Tao, and Min Zhang. TransGEC: Improving grammatical error correction with translationese. In **ACL**, 2023.

[16] Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. A simple recipe for multilingual grammatical error correction. In **ACL-IJCNLP**, 2021.

[17] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. LM-critic: Language models for unsupervised grammatical error correction. In **EMNLP**, 2021.

[18] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **ACL**, 2020.

[19] Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In **ACL**, 2020.

[20] Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. An empirical study of incorporating pseudo data into grammatical error correction. In **EMNLP-IJCNLP)**, 2019.

[21] Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In **BEA**, 2019.

[22] Maksym Tarnavskyi, Artem Chernodub, and Kostiantyn Omelianchuk. Ensembling and knowledge distilling of large sequence taggers for grammatical error correction. In **ACL**, 2022.

[23] Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. GECToR – grammatical error correction: Tag, not rewrite. In **BEA**, 2020.

[24] Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. Parallel iterative edit models for local sequence transduction. In **EMNLP-IJCNLP**, 2019.

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

[26] Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. **TACL**, 2016.

[27] Christopher Bryant, Mariano Felice, and Ted Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In **ACL**, 2017.

[28] Daniel Dahlmeier and Hwee Tou Ng. Better evaluation for grammatical error correction. In **ACL**, 2012.

[29] Takumi Gotou, Ryo Nagata, Masato Mita, and Kazuaki Hanawa. Taking the correction difficulty into account in grammatical error correction evaluation. In **COLING**, 2020.

[30] Md Asadul Islam and Enrico Magnani. Is this the end of the gold standard? a straightforward reference-less grammatical error correction metric. In **EMNLP**, 2021.

[31] Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. IMPARA: Impact-based metric for GEC using parallel data. In **COLING**, 2022.

[32] Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In **IJCNLP**, 2017.

A　各データセットの人手ランキング

表 5　レーティングアルゴリズムである TrueSkill (TS) を用いた各データセットの人手ランキング．

#	スコア	順位範囲	システム
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

GJG15 における文ベース評価

#	スコア	順位範囲	システム
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
8	-0.992	15	INPUT

SEEDA における文ベース評価

#	スコア	順位範囲	システム
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
	-0.231	13-14	BART
8	-0.797	15	INPUT

SEEDA における編集ベース評価

B　評価用プロンプト

LLM での編集ベース評価と文ベース評価で使用したプロンプトを図 2 と図 3 にそれぞれ示す．# context における [SOURCE] は原文であり，[PREVIOUS] と [FOLLOWING] はその原文の前後の文である．# targets における [CORRECTION N WITH EDITS] は編集が明示された訂正文（表 1 の訂正文を参考）であり，[CORRECTION N] は通常の訂正文である．ここで，N は 1 から 5 の値をとる．また，プロンプトは出力の形式を一定にさせるために，output format として JSON 形式で各スコアを出力する．観点ごとの評価では以下の文をプロンプトの第一段落の末尾に追加する．

- Difficulty: “Please evaluate each edit in the target with a focus on the difficulty of corrections.”
- Impact: “Please evaluate each edit in the target with a focus on its impact on the sentence.”
- Grammaticality: “Please evaluate each target with a focus on the grammaticality of the sentence.”
- Fluency: “Please evaluate each target with a focus on the fluency of the sentence.”
- Meaning Preservation: “Please evaluate each target with a focus on preserving the meaning between each target and the source, which is the middle sentence in the context.”

````
The goal of this task is to rank the presented targets based on the quality of each edit.
The context consists of three sentences from an essay written by an English learner.
After reading the context to understand the flow, please assign a score from a minimum of 1 point to a maximum of 5 points to each target based on the quality of the edit alone (note that you can assign the same score multiple times).
For targets without any edits, if the sentence is correct, they will be awarded 5 points; if there is an error, they will receive 1 point.
The edits in each target are indicated as follows:
Insert "the": [→the]
Delete "the": [the→]
Replace "the" with "a": [the→a]

# context
[PREVIOUS]
[SOURCE]
[FOLLOWING]

# targets
[CORRECTION 1 WITH EDITS]
        …
[CORRECTION N WITH EDITS]

# output format
The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "```json" and "```":

```json
{
   "target1_score": int  // assigned score for target 1
                  …
   "targetN_score": int  // assigned score for target N
}
```
````

図 2　編集ベース評価のプロンプト

````
The goal of this task is to rank the presented targets based on the quality of the sentences.
The context consists of three sentences from an essay written by an English learner.
After reading the context to understand the flow, please assign a score from a minimum of 1 point to a maximum of 5 points to each target based on the quality of the sentence (note that you can assign the same score multiple times).

# context
[PREVIOUS]
[SOURCE]
[FOLLOWING]

# targets
[CORRECTION 1]
        …
[CORRECTION N]

# output format
The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "```json" and "```":

```json
{
   "target1_score": int  // assigned score for target 1
                  …
   "targetN_score": int  // assigned score for target N
}
```
````

図 3　文ベース評価のプロンプト