文法誤り訂正における人手評価と自動評価の乖離とその解決

五藤 巧　坂井 優介　渡辺 太郎
奈良先端科学技術大学院大学
{goto.takumi.gv7, sakai.yusuke.sr9, taro}@is.naist.jp

概要

文法誤り訂正における自動評価尺度の目的の一つは，人手評価を模倣するような訂正システムの順位づけである．しかし，現状の自動評価は人手評価と乖離した評価手順に基づいており，このことは人手評価を模倣する目的と矛盾している．具体的には，人手評価は文単位の相対的な評価結果をレーティングアルゴリズムで順位に変換するが，自動評価では文単位の絶対的な評価結果を平均するなどしてコーパス単位評価に集約し，ソートすることで順位とする．本研究では，この乖離を埋めるために既存の自動評価尺度を人手評価の方法と一致するように用いることを提案し，実際に多くの尺度で人手評価との一致度が向上することを示す．

1　はじめに

文法誤り訂正分野は入力文に含まれる文法的な誤り，もしくは綴り・正書法の誤りなどを自動的に訂正することを目指す分野で，系列変換モデル [1, 2]，系列ラベリング [3, 4]，言語モデル [5, 6] による訂正システムが提案されている．ユーザは多様な訂正手法の中からできるだけ質の高い訂正システムを選択したいため，自動評価の結果に基づきシステムを順位付けすることで最適なシステムを検討する．自動評価には人手評価を模倣するような順位付けを行うことが期待されており，実際にこれまでの自動評価尺度のメタ評価が人手評価との一致を計ることにより実施されてきたことからも明らかである [7, 8]．例えば，自動評価の順位と人手評価の順位の間で Spearman の順位相関係数を計算し，より高い値を達成する尺度をよい尺度であるとみなすことができる．

しかし，人手評価の模倣を目的としているにも関わらず，現状の自動評価は人手評価とは乖離した手順に基づいている．図 1 に，2 文が含まれるデータセットを用いて 3 つの訂正モデルの順位づけを目的としたときの評価手順を示す．人手評価では，同じ入力文に対する訂正文をシステム間で相対的に比較し，比較結果を TrueSkill [9] などのレーティングアルゴリズムによって順位に変換する．一方，自動評価では，それぞれの訂正文を絶対的に評価することで評価値を推定し，コーパスレベルで平均もしくは累積した後ソートする．このように現状の自動評価は人手評価と乖離した評価手順を採用しており，人手評価を模倣するという目的と矛盾している．直感的には，自動評価は人手評価と同じ手順に基づいて実施することが望ましいと考えられる．

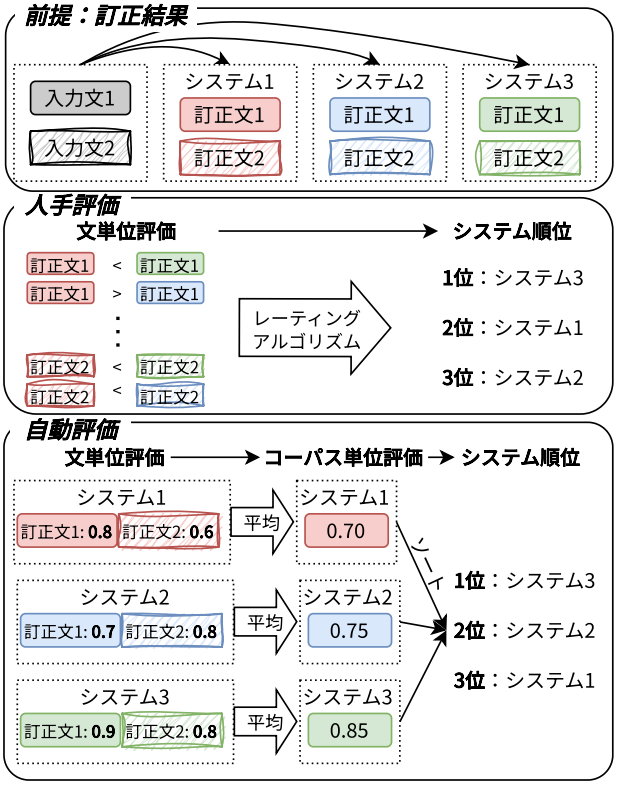


図 1: 2 文のデータセットに基づき 3 システムを順位づけるときの，人手評価および自動評価の概念図．

本研究では，この乖離を解決することで人手評価をよりよく模倣する自動評価が可能になるという仮説のもと，自動評価においても人手評価と同じ手

順で順位を計算することを提案する．実験では，既存の様々な自動評価尺度を対象に，メタ評価ベンチマークである SEEDA データセット [8] においてメタ評価を実施した．結果から，指摘した乖離を埋めることによって多くの尺度で評価結果が改善し，BERT [10] ベースの自動評価尺度でも大規模言語モデルを上回る評価が可能であることを述べる．また，今後の自動評価尺度の使用と開発について議論し，特に評価尺度の開発については文単位の相対的な評価性能が重要となることを述べる．

2 人手評価と自動評価の乖離

図 1 に示したように，人手評価と自動評価にはシステム順位の計算方法に乖離がある．人手評価は，CoNLL-2014 共通タスクに提出されたシステムを人手評価した Grundkiewicz ら [7] や，大規模言語モデルなどの最先端の訂正システムも含めて人手評価した Kobayashi ら [8] によって実施されてきた．いずれの研究でも，文単位における相対的な比較結果をレーティングアルゴリズムによりシステム順位に変換する．レーティングアルゴリズムには Expected Wins や TrueSkill [9] がよく用いられ，Grundkiewicz らは Expected Wins による結果を最終順位としており [7]，Kobayashi ら [8] は TrueSkill による結果を最終順位としている．

自動評価は，参照あり・なしや文単位・編集単位の尺度など多様な評価尺度を用いて実施されており，多くの尺度は各文の評価値を実数値で評価したあと，それらをコーパス単位の評価値に集約する手順を採用している．例えば，SOME [11] や IMPARA [12] などの文単位尺度では平均で集約し，ERRANT [13, 14] や GoToScorer などの編集ベース尺度，および GLEU [15, 16] や GREEN [17] などの n グラムベース尺度では，編集もしくは n グラムの数を累積して集約する．これらの方法で計算したコーパス単位の評価値は，ソートすることでシステム順位に変換することができる．これらの評価方法には，人手評価に見られるレーティングアルゴリズムに基づく処理は含まれない．ただし，Kobayashi ら [18] が提案した大規模言語モデルに基づく評価尺度は，文単位で 5 段階評価させた結果を TrueSkill によってシステム順位に変換しており，唯一乖離がない尺度であると考えられる．

3 乖離の解決法

人手評価と同じように自動評価尺度を使用することで乖離を解決する．まず，既存の自動評価尺度は入力文の評価値を実数値として計算するため，これらの値の大きさを比較することによって相対評価に変換する．図 1 の例では，入力文 1 に対する訂正文の評価値 0.8， 0.7， 0.9 を比較することで，人手評価で実施されるような比較結果に変換できる．その後，変換した相対評価の結果に TrueSkill を適用することで，システムの順位を計算する．本研究では，対象とするシステムから 2 つのシステムを取り出してペアワイズ評価することを考え，全ての組み合わせの結果を用いて TrueSkill を適用する．すなわち，システムの数を N とすると一文あたり $N(N-1)$ 回の比較結果が得られ，各文におけるこれらの結果からシステムの順位が計算される．

4 実験

4.1 自動評価尺度

各尺度のより詳細な実験設定は付録 A に示す．実装として gec-metrics[1] を開発して用いた．

ERRANT [13, 14] 訂正システムの編集と人手の編集の一致率に基づく参照あり尺度である．$F_{0.5}$ を使用することとし，参照が複数の場合は各文について $F_{0.5}$ が最も高くなる参照を選択する．

PT-ERRANT [19] 編集を BERTScore などのニューラルベースの尺度を用いて重みづける参照あり尺度である．ERRANT と同様，各文について $F_{0.5}$ が最も高くなる参照を選択する．

GLEU+ [15, 16] n グラムの一致に基づく参照あり尺度である．参照が複数の場合，各参照に対する評価値の平均を文単位の評価値とする．

GREEN [17] n-gram の一致に基づく参照あり尺度で，GLEU が precision に基づくこととは異なり F_β スコアを計算することができる．参照が複数の場合，各参照に対する評価値の平均を文単位の評価値とする．

SOME [11] 文法性・流暢性・意味保存性の観点でそれぞれ評価値を推定し，重み付き和を最終的な評価値とする参照なし尺度である．

IMPARA [12] 入力文と訂正文の意味保存性スコアと訂正文の品質推定スコアを用いる参照なし尺

1） https://github.com/gotutiyan/gec-metrics

表 1: SEEDA データセットを用いた人手評価との相関．*w/o TrueSkill* は従来の評価手順，*w/ TrueSkill* は提案法の評価手順で，従来の評価手順から改善したものに下線を引いた．太字は各列の最高値を示す．

	SEEDA-S				SEEDA-E			
	Base		+Fluency		Base		+Fluency	
Metrics	r (Pearson)	ρ (Spearman)	r	ρ	r	ρ	r	ρ
大規模言語モデルによる評価，Kobayashi ら [8] の論文値.								
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
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	0.763	0.706	-0.463	0.095	0.881	0.895	-0.374	0.231
PTERRANT	0.870	0.797	-0.366	0.182	0.924	0.951	-0.288	0.279
GLEU	0.863	0.846	0.017	0.393	0.909	0.965	0.102	0.486
GREEN	0.855	0.846	-0.214	0.327	0.912	0.965	-0.135	0.420
SOME	0.932	0.881	0.971	0.925	0.893	0.944	0.965	0.965
IMPARA	0.939	**0.923**	**0.975**	**0.952**	0.901	0.944	0.969	0.965
Scribendi	0.674	0.762	0.745	0.859	0.837	0.888	0.826	0.912

度である．

Scribendi [20]　言語モデルのパープレキシティと表層的な一致率を用いる参照なし尺度である．

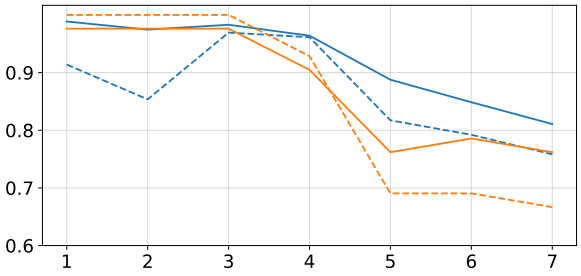
4.2　メタ評価方法

メタ評価のためのデータセットとして SEEDA [8] を用いる．文レベルで人手評価した SEEDA-S と編集レベルで人手評価した SEEDA-E の両方について，TrueSkill による人手評価結果に基づいてメタ評価結果を報告する．また，より大きく書き換えることも許した参照文および GPT-3.5 の出力を除いた Base 設定と，それらを含めた+Fluency 設定についても両方報告する．

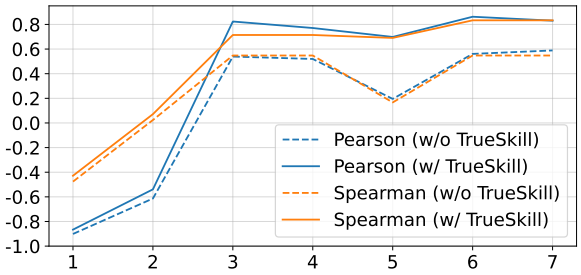
また，SEEDA で提案された分析手法の一つである window analysis により，計算された順位の頑健性を分析する．window-analysis では，システムを人手評価結果でソートした後，連続する N システムのみを対象に相関係数を計算する．これにより，人間から見て性能が拮抗するシステム集合を自動評価が正しく評価できるかを分析することができる．本研究では SEEDA の+Fluency 設定に対応する 14 システムを対象に $N = 8$ で実行し，Pearson と Spearman の相関係数を両方報告する．すなわち，人手評価における 1 位から 8 位，2 位から 9 位，...，7 位から 14 位までの結果それぞれについて相関係数を計算する．

4.3　実験結果

表 1 にメタ評価の結果を示す．中央のグループが従来の方法に基づく評価結果，下のグループが TrueSkill を使用することで人手評価と同じ方法で評価した結果である．上のグループは，Kobayashi ら [18] の GPT-4 による評価の結果で，SoTA に相当する．全体的な傾向として，TrueSkill に基づく評価によって多くの尺度で相関係数が向上している．特に，SEEDA-S および+Fluency における IMPARA の結果は，Kobayashi らの結果を上回った．また，ERRANT は多くの設定で 0.2 ポイント以上の改善が見られた．この結果から，人手評価と同じ評価手順で自動評価尺度を使用することで，人手評価とより近い順位付けが可能になることがわかる．言い換えると，これまで人手評価とは乖離した方法で自動評価尺度を使用していたことで，尺度の評価性能を低く見積もっていた．一方，GLEU や GREEN といった n グラムに基づく尺度には効果が見られなかった．人手評価は n グラムに基づいて実施されていないため，評価手順を一致させることでむしろ評価の



(a) IMPARA



(b) ERRANT

図 2: $N = 8$ における window-analysis の結果．横軸は人手順位の開始順位であり，例えば $x = 2$ は人手評価の 2 位から 10 位までのシステムを対象とした結果を示す．

粒度の違いが評価結果に悪影響を及ぼすと考えられる．Kobayashi ら [8] は自動評価と人手評価で評価の粒度を揃えることの重要性を指摘しており，この指摘は n グラムに基づく尺度の結果と整合する．

window-analysis の結果を図 2 に示す．IMPARA の結果は SEEDA-S に基づき，ERRANT の結果は SEEDA-E に基づいている．図 2a から，IMPARA では特に下位の順位について人手評価と一致しており，システム全体の相関の値も向上した．Pearson の積率相関係数では上位のシステムの評価結果も大きく改善した．また図 2b から，ERRANT では提案法により一貫して相関係数が改善したが，依然として上位のシステムの評価には苦戦している．上位のシステムには GPT-3.5 などの大きく書き換えることで訂正するシステムが含まれており，人手評価と計算方法を揃えたとしても，そのようなシステムの評価は依然として難しいと考えられる．

5 今後の自動評価尺度の使用と開発

使用について 既存の多くの尺度は一文ずつ評価値を推定する絶対評価尺度として使用されてきたが，今後は本稿のように複数の文を相対的に評価するように使用し，その結果をレーティングアルゴリズムでシステム順位に変換することを推奨する．ただし，いかなる状況でもこの評価手順が適切であるとは限らず，あくまでも “人手評価を模倣するシステム順位付けのための自動評価” という文脈に限定される議論である．例えば，人手評価の模倣ではなくシステムの弱点を分析する目的では，従来通り編集の数を累積する評価法のほうが解釈性が高く有用な可能性がある．また，今後人手評価の方法論が変化した場合には，自動評価の手順も人手評価の手順に追随するべきである．

開発について 本稿のように人手評価と自動評価の手順が一致している場合，文単位の性能比較において自動評価が人手評価を模倣するならば，システム順位も自動的に模倣する．したがって，今後は文単位の相対的な性能比較を高精度に行える尺度の開発に注力するべきである．同様のことは，表 1 の結果において SOME よりも IMPARA のほうが高い相関を達成することからも言える．これは品質推定器の学習時の観点が，IMPARA は相対評価，SOME は絶対評価に基づくことから説明できると考えられる．具体的には，IMPARA は多様な訂正文およびそれらの擬似評価値を impact という形で付与したデータセットを用いて訂正文の相対的な優劣を学習させる一方で，SOME は実数値で付与された人手評価値を推定するように学習される．この違いから，IMPARA のほうが相対的な評価に必要な情報を獲得できており，より人手評価を模倣するシステム順位を推定できたと考えられる．他にも，表 1 で高い相関を達成した Kobayashi ら [8] の尺度は，複数の訂正文を同時に入力することで相対的な観点での評価を可能とする尺度である．これらの結果から，自動評価において文単位の相対的な性能比較が重要であることがわかる．

6 おわりに

本研究では，文法誤り訂正における自動評価と人手評価との評価手順の乖離を埋めることを提案し，これにより人手評価とより一致する自動評価が可能になることを示した．具体的には，これまで絶対的な評価に用いられてきた尺度を相対的な評価に変換し，TrueSkill によりシステム順位を計算することで乖離を埋めた．また，今後の自動評価尺度を使用と開発の観点で展望し，特に開発については文単位の相対的な性能比較の重要性を述べた．

参考文献

[1] Satoru Katsumata and Mamoru Komachi. Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, **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**, pp. 827–832, Suzhou, China, December 2020. Association for Computational Linguistics.

[2] Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. A simple recipe for multilingual grammatical error correction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **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)**, pp. 702–707, Online, August 2021. Association for Computational Linguistics.

[3] Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. Parallel iterative edit models for local sequence transduction. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 4260–4270, Hong Kong, China, November 2019. Association for Computational Linguistics.

[4] Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. GECToR – grammatical error correction: Tag, not rewrite. In Jill Burstein, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch, editors, **Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 163–170, Seattle, WA, USA → Online, July 2020. Association for Computational Linguistics.

[5] Masahiro Kaneko and Naoaki Okazaki. Reducing sequence length by predicting edit spans with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 10017–10029, Singapore, December 2023. Association for Computational Linguistics.

[6] Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 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)**, pp. 205–219, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. Human evaluation of grammatical error correction systems. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 461–470, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[8] Masamune Kobayashi, Masato Mita, and Mamoru Komachi. Revisiting meta-evaluation for grammatical error correction. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 837–855, 07 2024.

[9] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: A bayesian skill rating system. In B. Schölkopf, J. Platt, and T. Hoffman, editors, **Advances in Neural Information Processing Systems**, Vol. 19. MIT Press, 2006.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. SOME: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 6516–6522, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

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

[13] Mariano Felice, Christopher Bryant, and Ted Briscoe. 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**, pp. 825–835, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[14] Christopher Bryant, Mariano Felice, and Ted Briscoe. 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)**, pp. 793–805, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[15] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. Ground truth for grammatical error correction metrics. In Chengqing Zong and Michael Strube, editors, **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)**, pp. 588–593, Beijing, China, July 2015. Association for Computational Linguistics.

[16] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. Gleu without tuning, 2016.

[17] Shota Koyama, Ryo Nagata, Hiroya Takamura, and Naoaki Okazaki. n-gram F-score for evaluating grammatical error correction. In Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito, editors, **Proceedings of the 17th International Natural Language Generation Conference**, pp. 303–313, Tokyo, Japan, September 2024. Association for Computational Linguistics.

[18] Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 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)**, pp. 68–77, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[19] Peiyuan Gong, Xuebo Liu, Heyan Huang, and Min Zhang. Revisiting grammatical error correction evaluation and beyond. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 6891–6902, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[20] Md Asadul Islam and Enrico Magnani. Is this the end of the gold standard? a straightforward reference-less grammatical error correction metric. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 3009–3015, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

A　評価尺度の詳細な実験設定

ERRANT について，Python モジュールの `errant=3.0.0` を用いて Span-based Correction の設定で評価した．参照の編集は人手で付与されているが，これを ERRANT により再抽出して用いた．PT-ERRANT は，重みづけを `bert-base-uncased` をモデルとした BERTScore により行い，F1 スコアに基づき重みを計算した．この時，ベースラインによる rescale を行い，idf による調整は行わない．ERRANT と同様，参照の編集は ERRANT により再抽出した．GLEU は，最大 4 グラムまでを用いて，イテレーション数を 500 とした．参照文のサンプルにおいてシード値は公式実装の設定に従った．GREEN は，最大 4 グラムまでを用いて，$F_{2.0}$ を評価値として用いた．SOME は，公式の学習済みモデルを用いて，文法性・流暢性・意味保存性の重みを公式実装に従い 0.43, 0.55, 0.02 にそれぞれ設定した．IMPARA は，品質推定モデルが公開されていないため再現実装・実験を行った．原論文に従い，CoNLL-2013 をシードコーパスとして 4096 件の学習ペアデータを生成し，これを 8:1:1 に分けた．8 に相当するものを学習データ，1 に相当するものを開発データとして `bert-base-cased` を追加学習した．アーキテクチャは transformers ライブラリの `BertForSequenceClassification` に従い，CLS トークンに相当する表現を実数値に線形変換する．学習コードは https://github.com/gotutiyan/IMPARA, 学習済みモデルは https://huggingface.co/gotutiyan/IMPARA-QE にある．推論時は，意味類似度を推定するモデルに `bert-base-cased` を用いて，推定された値の閾値を 0.9 とした．Scribendi は，言語モデルとして GPT-2[2] を用いて，Levenshtein distance ratio と token sort ratio の最大値に関する閾値を 0.8 として推論した．

文法誤り訂正の人手評価における TrueSkill の学習については，Sakaguchi ら [a1] の方法に従うことが一般的である．本研究でも基本的な実験設定は Sakaguchi らに従ったが，対戦するとみなすシステムのサンプル方法に関する修正は廃止し，全てのシステムペアを対象にする点は異なる．これにより，厳密には自動評価の手順は人手評価の手順と同じにはならないが，サンプル処理を含む場合は自動評価の再現性に影響を及ぼす可能性があるため，本稿では再現性が担保される利点を優先した．この再現性の問題はシード値を固定することで解決するようにも思えるが，ユーザが必ず共通のシード値で実験するとは限らず，高い評価性能を達成するようなシード値を恣意的に選択する余地を残してしまう．このような詳細な実験設定に関する議論は，将来の課題として慎重に議論する予定である．

[a1] Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. Efficient elicitation of annotations for human evaluation of machine translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 1–11, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

2) `https://huggingface.co/openai-community/gpt2`