順位学習に基づく日本語文章校正の品質推定モデルの構築

北岡 佑一 [1]　荒木 亮 [1]　真嘉比 愛 [1]　梶原 智之 [2]
[1] ちゅらデータ株式会社　[2] 愛媛大学／大阪大学
{y.kitaoka, m.araki, a.makabi}@churadata.okinawa　kajiwara@cs.ehime-u.ac.jp

概要

本研究では、日本語文章校正のための品質推定モデルを開発し、それを活用して文章校正モデルを改善する。GLEU などの既存の評価指標は文単位での評価を想定しており、文章の校正に対しては必ずしも適していない。そこで我々は、日本語の文章校正に適した品質推定モデル（参照なし評価）を構築する。ただし、文章の絶対的な品質アノテーションは高度な専門性を要しコストが高いため、文章間の相対的な品質比較を採用し、順位学習の枠組みで品質推定モデルを訓練する。評価実験の結果、2 つの日本語文章の優劣を判定する 2 値分類タスクにおいて、提案手法はドメイン内で 85%、ドメイン外でも 80%の正解率を達成し、GLEU などの既存の評価指標を上回る性能を示した。さらに、我々の品質推定モデルに基づく直接選好最適化によって、大規模言語モデルに基づく文章校正の性能を改善できた。

1　はじめに

大規模言語モデルの登場により、文章校正の自動化技術 [1] も飛躍的な進歩を遂げている。しかし、大規模言語モデルによるハルシネーションや過剰編集などの課題 [2] も依然として残っている。これらの課題に対処し、文章校正モデルを継続的に改善していくためには、人手評価との高い相関を持つ自動評価の確立が重要である。特に、我々が開発および運用している文章校正サービス「ちゅらいと」[1)] のような実環境においては、ユーザからの入力に対して参照テキストは存在しないため、参照なし評価である品質推定の技術が望まれる。

文法誤り訂正を含む文章校正の先行研究では、参照ベース自動評価の GLEU [3] や GREEN [4]、参照なし品質推定の SOME [5] や IMPARA [6] などの評価指標が提案されてきた。しかし、これらの評価指標は文単位での評価を想定しており、表記揺れやスタイルなどの文を超える一貫性や、改行や箇条書きなどの文書構造まで考慮する必要がある文章全体の校正に対しては、必ずしも適していない。

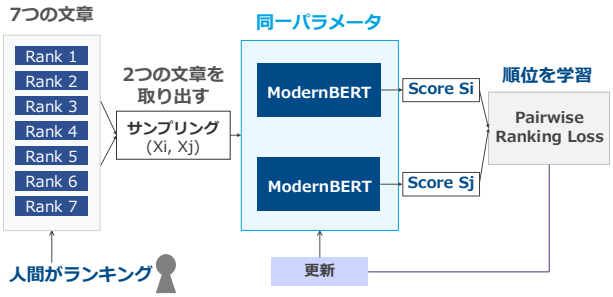

図 1　提案手法の概要

本研究では、日本語の文章校正を対象に、文章単位での品質推定に取り組む。ここで、品質推定モデルの訓練には、テキストに対して品質ラベルを付与したデータセットが必要である。しかし、文章の絶対的な品質アノテーションは、高度な専門性を要するタスクでありコストが高い。そこで我々は、文章間の相対的な品質比較に基づく順位学習のアプローチを採用し、品質推定モデルを訓練する。さらに、本研究で構築する品質推定モデルを用いて擬似ラベル付きデータセットを作り、自己学習に活用する。

日本語の Web テキストを対象とする実験の結果、2 つの文章の優劣を判定する 2 値分類タスクにおいて、提案手法は 85%の正解率を達成し、GLEU などの既存の参照ベース自動評価の性能を上回った。また、ドメイン外の Wikipedia を対象とする実験においても、同様の設定で 80%の正解率を達成し、高い汎化性能を示した。さらに、本研究で構築した品質推定モデルを直接選好最適化 (Direct Preference Optimization; DPO) [7] に基づく自己学習 [8] に応用したところ、大規模言語モデルに基づく文章校正の性能を改善できることが明らかになった。提案手法は、文章の相対的な品質比較のみを必要とする低コストな方法で、文章校正モデルを改善できる。

1)　https://chuwrite.com/

2 提案手法

本研究では、文章校正の品質推定のために、文章 x に対して品質スコア $s \in \mathbb{R}$ を推定する回帰モデル $s = f_\theta(x)$ を訓練する。ここで、θ は Transformer ベース回帰モデル [9] のパラメータである。なお、品質スコア s が大きいほど、望ましい文章と考える。

2.1 アノテーション

品質推定モデルの訓練には、文章と品質ラベルから成るデータセットが必要である。本研究では、所与の文章に対して 3.1 節で紹介する複数の校正モデルから校正候補を収集し、それらをランキングして品質ラベルを付与する。ただし、同等の品質を持つ文章間には、同ランクを付与することを許す。

絶対的な品質スコア [10] を付与する方法も考えられるが、絶対評価は評価者の主観に依存するため、評価者間でのスケール差や評価基準の曖昧性などの課題がある。一方で、本研究で採用する相対評価は評価者にとってより直感的であり、品質の高低に依らず候補の優劣を評価しやすいという利点を持つ。

2.2 順位学習

前節で構築したデータセットにおける相対評価のラベルを有効活用し、ペアワイズ順位学習 [11] によって品質推定モデルを訓練する。図 1 に示すように、データセット $\mathcal{D}$ から文章ペア (x_i, x_j) とその品質ラベル $y_{ij} \in \{-1, 0, 1\}$[2] を抽出し、2 つの品質推定モデルを通して品質スコア s_i および s_j を得る。これらの 2 つの品質推定モデルはパラメータ θ を共有し、以下の margin-based pairwise ranking loss[3] で訓練する。ただし、N はバッチサイズである。

$$\mathcal{L} = \frac{1}{N} \sum_{(x_i, x_j, y_{ij}) \in \mathcal{D}} \ell(s_i, s_j, y_{ij}) \tag{1}$$

$$\ell(s_i, s_j, y_{ij}) = \begin{cases} \max(0, \gamma - y_{ij}(s_i - s_j)) & \text{if } y \neq 0 \\ |s_i - s_j| & \text{if } y = 0 \end{cases} \tag{2}$$

これは、品質の順序関係を満たしつつ、品質スコアの差がマージン γ 以上となるように品質推定モデルを訓練する [12] という考え方である。ただし、同ランクの文章間では、品質スコアの差を最小化する。

2) 品質ラベルは、x_i の品質が x_j の品質よりも高い場合に $y_{ij} = 1$、低い場合に $y_{ij} = -1$、同等の場合に $y_{ij} = 0$ とする。

3) PyTorch の MarginRankingLoss を、同ランクにも対応できるように拡張した。 https://docs.pytorch.org/docs/stable/generated/torch.nn.MarginRankingLoss.html

表 1 データ構築用の校正モデルと人手評価の平均順位

モデル	追加訓練	サイズ	平均順位
正解文章	-	-	1.00
入力文章	-	-	2.15
GPT-4o	-	Unspecified	2.17
ちゅらいと	-	Unspecified	2.51
GPT-OSS	-	20B	2.91
T5	JWTD	3B	5.24
Sarashina	JWTD	13B	5.43

3 内的評価：品質推定のメタ評価

品質推定モデルを訓練し、そのメタ評価のために、2 つの文章の優劣を判定する 2 値分類を行う。そして、品質推定と人手評価の一致率を評価する。

3.1 品質推定モデルの構築

データセットの構築 日本語の Web テキスト 104 記事を人手で校正したうえで、品質アノテーションに使用した。これらの元記事は、平均約 2,500 文字の長文である。表 1 に示す 5 つの校正モデルによる校正候補に入力文章および正解文章を加えた 7 つの文章を 1 組として、ランキングによる品質アノテーション[4] を実施した。ここで使用したモデルは、文章校正サービスちゅらいと[5]、大規模言語モデルの GPT-4o[6] [13] および GPT-OSS[7] [14]、JWTD[8] [15] 上で訓練した誤り訂正モデルのうち予備実験で高い性能を示した T5[9] [16] および Sarashina[10] である。

人手評価によるランキングの結果を表 1 に示す。JWTD で追加訓練したモデルは低評価となる場合が多かった。これは、JWTD が文単位の編集を対象としているため、文章全体での一貫性や文書構造に関する校正に十分に対応できなかったためだと考えられる。追加訓練なしの大規模言語モデルは追加訓練モデルよりは高評価であったが、改悪リスクのない入力文章そのままと同等以下の平均順位であった。

品質推定モデルの訓練 上記のランキングから、${}_7C_2 = 21$ 通りずつのペアデータを作成し、品質推定

4) 評価者は、日本語母語話者である社内のアノテータ 5 名。

5) ちゅらいとは本来、ユーザが校正規則を選択できる校正支援ツールである。本実験では全規則を適用したため、実サービスほどの性能を発揮できていないことには注意されたい。

6) https://platform.openai.com/docs/models/gpt-4o

7) https://huggingface.co/openai/gpt-oss-20b

8) https://nlp.ist.i.kyoto-u.ac.jp/?日本語 Wikipedia 入力誤りデータセット

9) https://huggingface.co/retrieva-jp/t5-xl

10) https://huggingface.co/sbintuitions/sarashina2-13b

表 2　メタ評価：人手評価との一致率

評価指標	参照ベース	Web 記事	JWTD
GLEU	✓	0.600	–
GREEN	✓	0.650	–
提案手法		**0.850**	**0.799**

モデルの訓練に使用した。データセットは、訓練用 95 記事（1,995 ペア）、検証用 5 記事（105 ペア）、評価用 4 記事（84 ペア）に分割して使用した。

品質推定モデルは ModernBERT[11] [17] を 2.2 節の手法でファインチューニングして構築した。ただし、損失関数のマージンは $\gamma = 0.5$ とした。最適化には AdamW [18] を使用し、バッチサイズは 32、最大エポック数は 20 に設定した。学習率は 5×10^{-6} とし、学習率のスケジューリングには Linear Warmup (warmup ratio 0.1) を採用、Weight Decay は 0.01 に設定した。推論には vLLM[12] を用いた。

3.2　実験設定

品質推定モデルのメタ評価のために、品質推定と人手評価の一致率を評価した。比較手法として、参照ベース自動評価の GLEU [3] および GREEN [4] を採用し、gec-metrics[13] [19] の実装を用いて実験した。ここで、比較手法の GLEU および GREEN は、入力文章および正解文章を参照しつつ校正候補を評価する。そのため、本実験では 7 つ組データセットのうち入力文章および正解文章を除外し、${}_5C_2 \times 4$ 記事 = 40 ペアを評価に使用した。

3.3　実験結果

表 2 に実験結果を示す。提案手法は人手評価と 85%一致し、2 つの文章の優劣を判定するタスクにおいて優れた性能を示した。提案手法は GLEU などの既存の自動評価よりも高性能であるのに加えて、参照文章との比較を必要としない利点も持つ。

品質推定モデルの訓練データとは異なるドメインの JWTD（Gold データの 1,127 文対）においても、提案手法の有効性を検証した。JWTD における誤り文と訂正文の対を提案手法によって評価した結果、表 2 に示すように、80%の正解率で訂正文に高い評価値を付与できた。さらに、JWTD における失敗例を表 3 に示す。これらの例では、括弧内の左側が

11）https://huggingface.co/sbintuitions/modernbert-ja-310m
12）https://github.com/vllm-project/vllm
13）https://github.com/gotutiyan/gec-metrics

表 3　提案手法が誤り文を高く評価してしまった失敗例

ベテラン声優の中には副業を **{行なう → 行う}** 者もいる
東茨城郡茨城町下石崎 **{樋沼湖畔 → 涸沼湖畔}** キャンプ場
退役した機体は **{ギリシア → ギリシャ}** に引き渡されている

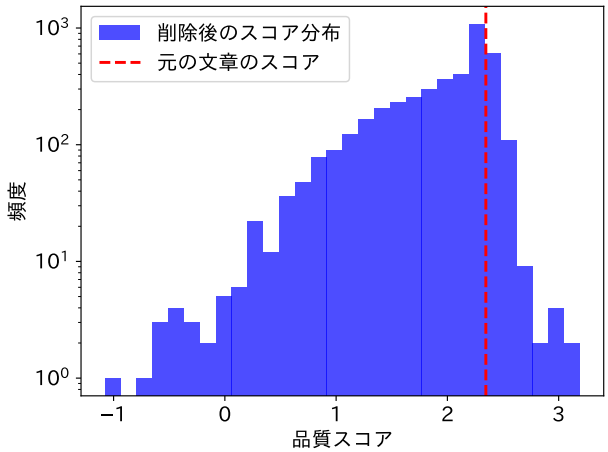

図 2　正解文章から 1 文字削除した際の品質スコアの分布

JWTD における誤り表現であり、右側が訂正表現である。提案手法の失敗例は、これらのように、送り仮名や翻字などの表記揺れや固有名詞の誤りに関するものであり、人間の校正者にとっても判断が難しい事例ばかりであった。これらの実験結果から、提案手法の高いドメイン汎化性能を確認できた。

3.4　分析：1 文字の編集に対する感度

本実験に使用した Web 記事のうち 1 件を使用し、人手で校正した正解文章から無作為に 1 文字ずつを削除した場合の提案手法の品質スコア分布を分析した。図 2 に品質スコアの分布を示す。多くの場合に期待通り品質スコアが低下したものの、約 1 割の事例では正解文章以上の品質スコアが見られた。

1 文字の削除によって品質スコアが向上した事例には、**五月雨登校 → 五月登校**のように、トークン数を減少させる編集が見られた。**五月_雨_登校**という 3 トークンの表現から “雨” を削除すると**五月_登校**という 2 トークンに変化する。このように、意味変化を伴いつつも、トークン数が減少する事例では、品質スコアの向上が見られた。一方で、**せっかくだから → せ_っく_だから**のように、“か” の 1 文字を削除することでトークン数が増加するような事例では、品質スコアが低下する傾向があった。このような、サブワード分割による品質スコアへのバイアスは、提案手法における課題のひとつである。

4　外的評価：自己学習への応用

品質推定を文章校正モデルの訓練にも応用する。7,301 件のニュースを対象に、GPT-OSS[7] [14] に基づく文章校正モデルを、DPO [7] ベースの自己学習 [8] で改善する。この自己学習[14]では、GPT-OSS による校正前後の文章対に対して、品質推定によって正例と負例を定義し、正例の出現確率を高める。

具体的には、GPT-OSS (π_{old}) が生成した校正候補 $\hat{y}$ および入力文章 x を品質推定モデル S で比較し、スコアが高い方を正例 y^+、低い方を負例 y^- とする。ここで、DPO による更新後の GPT-OSS を π_{new}、更新前の GPT-OSS を π_{old}、$\sigma(\cdot)$ をシグモイド関数、β を温度パラメータとして、DPO ベース自己学習の損失関数は以下のように記述できる。

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y^+,y^-)}\left[\log \sigma\left(\beta\Delta\right)\right], \tag{3}$$

$$\Delta = \left(\log \frac{\pi_{\text{new}}(y^+|x)}{\pi_{\text{old}}(y^+|x)} - \log \frac{\pi_{\text{new}}(y^-|x)}{\pi_{\text{old}}(y^-|x)}\right). \tag{4}$$

4.1　品質推定に基づくフィルタリング

予備実験として、校正前後の文章対のうち品質推定によって高評価を得た文章を正例とする DPO ベース自己学習を実施した。訓練用と検証用で 9:1 にデータセットを分割し、検証用データにおいて選好正解率を評価したところ、60%前後の性能に留まった。この低性能は、僅差の品質を持つ文章対が学習のノイズとなることが原因だと考えられる。

予備実験の結果を踏まえて、品質推定のスコア差が一定以上の文章対のみを自己学習に用いることにした。具体的には、GPT-OSS の校正によって品質推定のスコアが 0.05 以上向上した文章対のみを用いた。DPO データセットのフィルタリングによって、自己学習に使用する訓練データ数は 6,570 件から 1,789 件に減少したが、選好正解率を改善できた。

4.2　定量評価

3.1 節の評価セットを用いて、DPO ベース自己学習の有効性を評価した。本実験では、自己学習を適用しない GPT-OSS、校正結果を常に正例とする品質推定なしの自己学習、品質推定に基づくフィルタリングを適用した自己学習（提案手法）の性能を比較した。評価指標には我々の品質推定[15]を用いた。

14）Unsloth を用いて訓練した https://unsloth.ai/
15）平均が 0、標準偏差が 1 となるように標準化を施した。

表 4　GPT-OSS に対する DPO ベース自己学習の効果

	平均順位	平均スコア
提案手法	1.91	0.12
自己学習なし	1.99	−0.08
品質推定なしの自己学習	2.10	−0.04

表 4 に実験結果を示す。品質推定なしの単純な DPO ベース自己学習では、自己学習を適用する前の GPT-OSS よりも校正品質が悪化した。一方で、品質推定に基づくフィルタリングを適用した提案手法では、自己学習によって校正品質を改善できた。特に、校正結果の品質スコアが、自己学習なしではマイナスだったものが、提案手法ではプラスに好転した。この実験から、我々の品質推定を活用して、文章校正モデルを改善できることが確認できた。

4.3　定性評価

自己学習を経た文章校正モデルの出力を定性的に分析すると、改行や箇条書きなどの文書構造の整理に関する校正が適切になされる傾向が見られた。これは、文章単位の品質推定の特性が、DPO を通じて文章校正モデルに適切に反映された結果であると考えられる。一方で、誤字や脱字などに関する局所的な校正能力は必ずしも十分に改善できているとは言えず、プロンプトに影響を受ける傾向が見られた。

5　まとめ

本研究では、日本語の文章校正のための品質推定（参照なし評価）モデルを構築した。これは、文章校正モデルの評価に有用なだけでなく、文章校正モデルの訓練にも活用が期待できる。内的評価では、品質推定モデルがドメイン内評価で 85%、ドメイン外評価においても 80%の性能で人手評価と一致することが示された。これは、GLEU [3] などの既存の参照ベース評価指標を上回る性能であり、文章校正の評価のための有用性を確認できた。また、外的評価では、DPO [7] ベースの自己学習 [8] において、学習データのフィルタリングのために品質推定を活用した。自己学習の結果、大規模言語モデルに基づく文章校正モデルの性能改善を確認できた。

今後の課題として、3.4 節および 4.3 節で述べた、誤字や脱字などの局所的な編集に対する評価性能の改善がある。また、リランキングやプロンプト最適化 [20] など、品質推定を応用した文章校正モデルの更なる改善にも取り組みたい。

参考文献

[1] Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. 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**, pp. 17–33, 2024.

[2] Xinyuan Li and Yunshi Lan. Large Language Models are Good Annotators for Type-aware Data Augmentation in Grammatical Error Correction. In **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 199–213, 2025.

[3] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 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**, pp. 588–593, 2015.

[4] Shota Koyama, Ryo Nagata, Hiroya Takamura, and Naoaki Okazaki. n-gram F-score for Evaluating Grammatical Error Correction. In **Proceedings of the 17th International Natural Language Generation Conference**, pp. 303–313, 2024.

[5] 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, 2020.

[6] 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, 2022.

[7] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In **Proceedings of the 37th Conference on Neural Information Processing Systems**, pp. 53728–53741, 2023.

[8] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-Rewarding Language Models. In **Proceedings of the 41st International Conference on Machine Learning**, pp. 57905–57923, 2024.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Proceedings of the 31st Conference on Neural Information Processing Systems**, pp. 5998–6008, 2017.

[10] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Can Machine Translation Systems Be Evaluated by the Crowd Alone. **Natural Language Engineering**, Vol. 23, No. 1, pp. 3–30, 2017.

[11] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to Rank using Gradient Descent. In **Proceedings of the 22nd International Conference on Machine Learning**, pp. 89–96, 2005.

[12] Thorsten Joachims. Optimizing Search Engines using Clickthrough Data. In **Proceedings of The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, pp. 133–142, 2002.

[13] OpenAI. GPT-4o System Card. **arXiv:2410.21276**, 2024.

[14] OpenAI. Gpt-oss-120b & Gpt-oss-20b Model Card. **arXiv:2508.10925**, 2025.

[15] Yu Tanaka, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. Building a Japanese Typo Dataset from Wikipedia's Revision History. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**, pp. 230–236, 2020.

[16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[17] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics**, pp. 2526–2547, 2025.

[18] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In **Proceedings of the Seventh International Conference on Learning Representations**, 2019.

[19] Takumi Goto, Yusuke Sakai, and Taro Watanabe. gecmetrics: A Unified Library for Grammatical Error Correction Evaluation. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics**, pp. 524–534, 2025.

[20] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large Language Models Are Human-Level Prompt Engineers. In **Proceedings of the Eleventh International Conference on Learning Representations**, 2023.