文法誤り訂正における参照なし評価尺度を用いた分析的評価法

五藤巧 [1] 渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学
{goto.takumi.gv7, taro}@is.naist.jp

概要

文法誤り訂正における参照なし評価尺度は様々な利点を持つにも関わらず，その評価値の向上もしくは低下がどの訂正に起因するものなのかを分析することが難しい．本研究では，協力ゲーム理論におけるシャープレイ値の方法に注目し，訂正前と訂正後の文単位の評価値の差分を，訂正単位の評価値に分配する方法を提案する．これにより，全体の評価値に対する個々の訂正の貢献度を可視化し，分析を可能にする．実験では，実際の訂正をニューラルベースの尺度で評価する時の分析例を示す．また提案法の妥当性や将来的な応用について議論する．

1 はじめに

文法誤り訂正は入力に含まれる文法的もしくは表層的な誤りを自動的に訂正するタスクである．同タスクの評価は参照あり評価と参照なし評価に大別され，特に後者は人手による訂正文を必要としない．このことから，低コストで評価可能な点が利点である．この他にも，仮にシステムの訂正が正しくても，参照あり評価では人手の訂正に含まれない訂正は誤りとなってしまうが，参照なし評価ではそのような問題は起こりづらいと考えられる．さらに，ある特定の観点から評価することで，単なる正誤に留まらない評価が可能である．例えば，Asano ら [1] や Yoshimura ら [2] は文法性・流暢性・意味保存性を対象とした観点別の評価を可能にしている．

こうした尺度のほとんどはニューラルベースであるが，これらの尺度は文単位の評価値のみを提供するため，評価値が向上もしくは低下した場合に，システムのどのような訂正によって評価値が向上（低下）したのかを分析することはできない．このことから，システムをどのように改良すればよいのかという知見が得られない．

本研究では，参照なし評価尺度を用いて，どのような訂正によって評価値が向上もしくは低下したのかを分析可能にするため，訂正前および訂正後の文の評価値の差分を，訂正単位の評価値に分配する方法を提案する．得られる訂正単位の評価値は実数値であり，正負が評価尺度にとっての善し悪しを表し，絶対値の大きさが善し悪しの強度を表す．分配方法には，協力ゲーム理論の文脈で発展してきたシャープレイ値 [3] を応用し，シャープレイ値が持つ理論的な性質が提案法にとってうまく解釈できることを議論する．提案法は，参照なし評価の利点を維持しながら，個々の訂正の質を分析する分析的評価法を提供するものである．

実験では，ニューラルベースの参照なし評価尺度として SOME [2] と IMPARA [4] を用いて提案法を適用し，どのような解釈が可能であるかを例示する（§ 4.3）．分析では，提案法の妥当性について議論し（§ 5.1），また提案法が分析的評価法以外の用途でも有効に機能することを議論することで，提案法の展望について述べる（§ 5.2）．具体的には，メタ評価法，脆弱性やバイアスを発見する手法，説明性手法の観点から議論する．

2 関連研究

参照なし評価尺度は，言語モデルのパープレキシティと表層の一致率に基づくもの [5]，流暢性・文法性・意味保存性の各観点に基づくもの [1, 2]，意味保存性推定と訂正文の品質推定に基づくもの [4] といったように，さまざまな手法が提案されている．しかしながら，いずれも評価値の変動がどの訂正に起因するものなのかという情報は提供できない．本研究ではこれらの評価法の一部を用いて実験する．

永田ら [6] は，書き手の意図が伝わるかどうかという観点から訂正重要度を定量化する方法を提案している．永田らは人間にとっての（意味保存性の観点からの）訂正重要度を定量化するが，提案法は参照なし評価尺度にとっての訂正重要度を定量化していると解釈できる．すなわち，人間の判断を模倣した評価値の獲得は目的としていない点が異なる．

3 分析的評価法

参照なし評価は，参照なし評価尺度 $M(\cdot)$ に誤り文 S とその訂正文 H を入力し，訂正文の評価値 $M(S,H) \in \mathbb{R}$ を与えるものである．この時，自動アライメント手法である ERRANT [7, 8] により，S を H にするための訂正 $\boldsymbol{e} = \{e_i\}_{i=1}^{N}$ を獲得できる．N は訂正の数であり，$e \in \boldsymbol{e}$ は S における単語単位のスパンと，そのスパンの訂正後の文字列を含む値である．

本研究の基本アイデアは，$\boldsymbol{e}$ を適用することで変化した文単位の評価値 $\Delta M(S,H) = M(S,H) - M(S,S)$ は，訂正 $\{e_i\}_{i=1}^{N}$ それぞれが持つ貢献 $\{\phi_i \in \mathbb{R}\}_{i=1}^{N}$ の総和であるという仮説に基づく．したがって，

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

となるような $\{\phi_i\}_{i=1}^{N}$ を求めることで，個別の訂正が持つ評価値への貢献を定量化でき，分析的評価に用いることができる．特に，$\phi_i > 0$ であれば尺度 $M(\cdot)$ を改善する訂正であり，$\phi_i < 0$ であれば悪化させる訂正であることが分かる．さらに，ϕ_i の絶対値の大きさによってその度合いも定量化できる．例えば，ϕ_i が大きな負の値であれば，尺度にとって改善されることが特に望ましい訂正であることが分かる．

次に，ϕ_i をどのように求めるかが課題となる．提案法では，協力ゲーム理論で用いられるシャープレイ値 [3] に注目する．複数のプレイヤーが協力してある利益を達成したとき，各プレイヤーの貢献度を考慮しながら全体の利益を配分したときの利益がシャープレイ値である．本研究では全体の利益を $\Delta M(S,H)$，プレイヤーを編集 $\boldsymbol{e}$ であるとみなして，尺度 $M(\cdot)$ に対するシャープレイ値 $\{\phi_i(M)\}_{i=1}^{N}$ を次式で計算する．

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

なお，訂正集合 $E \subseteq \boldsymbol{e}$ を入力文 S に適用した訂正文を S_E と表した．式 2 は，全体の訂正集合から e_i を含まない全ての部分集合について e_i を適用した時と適用しない時の評価値の差分を計算し，その加重和をとったものである．より具体的な計算例は付録 A に示す．

シャープレイ値は，次の 4 つの性質を満たす唯一の分配方法であるとして知られる [3]．以下，その性質を文法誤り訂正における解釈と共に説明する．

1. **効率性**：式 1 そのものである．訂正集合全体の貢献は，個々の訂正の貢献の和に一致する．
2. **対称性**：$E \subseteq \boldsymbol{e} \setminus \{e_i, e_j\}$ を満たす任意の訂正集合 E について，$\Delta M(S, S_{E \cup e_i}) = \Delta M(S, S_{E \cup e_j})$ ならば，$\phi_i(M) = \phi_j(M)$ である．すなわち，評価値への貢献が同じ度合いである訂正のシャープレイ値は一致する．
3. **ダミープレイヤ**: 任意の $E \subseteq \boldsymbol{e} \setminus e_i$ について，$\Delta M(S, S_{E \cup e_i}) = \Delta M(S, S_E)$ ならば，$\phi_i = 0$ である．すなわち，評価値に影響しない訂正の貢献は 0 となる．文法誤り訂正においては多くの単語は訂正されずに維持されるが，「訂正しない」という訂正が $\boldsymbol{e}$ に含まれると仮定しても，それらの訂正の貢献は $S_{E \cup e_i} = S_E$ より必ず 0 となる．このことから，実際に何らかの変更を加えた訂正のみに対して評価値が分配されることを保障する．
4. **加法性**：2 つの異なる評価尺度 M および M' を結合した評価尺度 $\Delta M(S,H) + \Delta M'(S,H)$ に対して，$\phi_i(M + M') = \phi_i(M) + \phi_i(M')$ が成り立つ．意味保存性と流暢性のように複数の観点を考慮した場合のシャープレイ値は，個々の観点におけるシャープレイ値の和として計算できることを意味する．

以上のように，シャープレイ値の持つ性質は，本研究で目的とする分析的評価法と相性が良い．また式 2 より，シャープレイ値はある訂正の貢献を計算するときに周辺の訂正の適用状況を考慮するため，訂正の依存関係も考慮した評価値が得られると考えられる．なお，厳密なシャープレイ値を計算する場合，式 2 を計算するために一文あたり $\mathcal{O}(2^N)$ の時間計算量を必要とし，また任意の $E \subseteq \boldsymbol{e}$ について $M(S, S_E)$ を計算するため 2^N 文に対する評価値の計算を必要とすることに注意されたい．

4 実験

4.1 参照なし評価尺度

SOME [2] 人手評価を直接教師として，BERT を文法性・流暢性・意味保存性のそれぞれの観点において学習する手法である．本研究ではこの 3 種類の

表 1　実際の訂正を用いた提案法の適用例.

原文	-	It	is	also	take	risks	raher	than	only	doing		.
訂正文	-	It		also	takes	risks		than	just	doing	things	.
尺度	$\Delta M(S,H)$	編集単位の評価値 ϕ_i										
SOME-f	0.3214	-	0.090	-	0.038	-	0.101	-	0.045	-	0.048	-
SOME-g	0.2427	-	0.064	-	0.024	-	0.075	-	0.042	-	0.037	-
SOME-m	0.2332	-	0.002	-	0.008	-	0.124	-	0.005	-	0.095	-
IMPARA	-0.0004	-	0.024	-	0.066	-	0.066	-	-0.155	-	-0.002	-

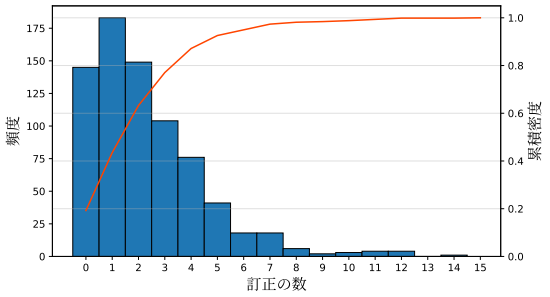

図 1　JFLEG 開発データの GECToR による訂正文における，訂正の数の頻度（棒グラフ）と頻度の累積密度（折れ線グラフ）.

観点の尺度に対して独立に提案法を適用し，観点別の分析的評価が可能であることを示す．なお，学習済み評価器の重みには公開されているものを用いた[1)].

IMPARA [4]　IMPARA は，意味保存性の評価値と訂正文の品質推定の評価値を組み合わせて評価する．評価器の重みは公開されていないため，再現実装を行い，CoNLL-2013 [9] を並列データとして学習したモデルを用いた[2)].

4.2　訂正文

提案法は分析的評価を目的としているため，実験に用いる訂正システムの訂正性能は基本的に問わない．本稿では，系列タグ付けとして文法誤りを訂正するモデルである GECToR [10] の出力を用いて提案法の結果を例示する．JFLEG [11] の開発データ 754 文を入力とし，推論時のハイパーパラメタ：KEEP タグへのバイアスと誤り検出確率の閾値は共に 0 として推論した．また，RoBERTa [12] を fine-tune することで学習された単一モデルを用いた．

得られた訂正文における訂正数の分布および累積密度を 1 に示す．3 節で述べたように提案法の時間計算量は指数関数のオーダであるが，図 1 より最大でも $N <= 14$ であるため，全ての文において現実的な時間でシャープレイ値が計算可能である．

4.3　分析的評価の例示

表 1 に，提案法の結果を例示する．表 1 は 5 つの誤り訂正が行われた実例である．原文と訂正文は，ERRANT [7, 8] が計算するアライメントにしたがってチャンクに分けたものを示した．最下部のブロックには，何らかの訂正が行われたチャンクそれぞれに対して，4.1 節で示した尺度を $M(\cdot)$ として提案法を適用した時のシャープレイ値を示した．SOME-f・SOME-g・SOME-m は，SOME の流暢性・文法性・意味保存性の評価結果をそれぞれ示す.

表 1 において $\Delta M(S,H)$ に注目すると，SOME-f・SOME-g・SOME-m は評価値が向上している一方で，IMPARA の評価値はわずかに低下している．各尺度においてこの理由を探るために，訂正単位の評価値に注目すると次のように分析できる．まず，SOME-f の評価値が向上した理由として，*raher* や *is* の削除の訂正によるものが大きいことが分かる．その他の訂正も正の評価であるが，その度合いは小さい．SOME-g の分析結果としても同じことが言える．SOME-m は，*raher* の削除と *things* を追加する訂正により主に評価値が向上している．IMPARA は，ほとんどの訂正が正の評価を与える一方で，副詞の語彙選択（*only* → *just*）の訂正が大きな負の評価値を与えており，文単位の評価値が低下する原因となっている．したがって，IMPARA の観点ではこの訂正が改善されることが望ましい訂正であることが分かる.

なお，表 1 は，それぞれの尺度や観点において分析的評価が可能なことの例示を目的としており，メタ評価のように尺度による評価の傾向の違いを議論するものではないことに注意されたい．また，提案法で得られた評価値の絶対値の大きさは，同じ尺度

1）https://github.com/kokeman/SOME
2）https://github.com/gotutiyan/IMPARA

の中で異なる訂正同士を比較してよいが，同じ訂正の評価値を異なる尺度で比較できないことに注意されたい．この理由は，評価値の分配元である $\Delta M(\cdot)$ の大きさによって，訂正単位の評価値大きさの解釈が変動するからである．

5 議論

5.1 分析的評価法としての妥当性

提案法により得られた訂正単位の評価値の妥当性には議論の余地がある．まず，分析的評価法としての妥当性の評価に人手評価を用いることは望ましくない．本研究では，文単位の評価値の向上もしくは低下の理由を，訂正単位の評価値によって分析することを分析的評価と位置付けており，ここに人手評価を模倣するような分析を可能にする意図は含まれていない．したがって，人間の直感と一致するかどうかに関わらず，得られた訂正単位の評価値をもとに分析することを前提としている．

一方で，依然として訂正単位の評価値の計算方法としてシャープレイ値が適切かどうかの議論は残る．訂正単位の評価値を定量化する単純な方法には，入力文にただ一つの訂正のみを適用した時の評価値の変動に基づく方法 [13]（S と S_{e_i} のみの比較）や，訂正文からただ一つの訂正のみを除外したときの評価値の変動に基づく方法 [4]（$S_{\boldsymbol{e}}$ と $S_{\boldsymbol{e}\setminus e_i}$ のみの比較）が考えられる．しかし，いずれも周辺の訂正の適用状況を考慮できない．これらの方法に比べて，シャープレイ値は考えうる全ての訂正の適用状況の組み合わせを考慮するため，周辺の訂正の適用状況を適切に考慮した評価値を得ることができる．このことは，メタ評価の一環として尺度の誤りタイプに対する敏感性を調査した Choshen ら [14] の方法を，より厳密な計算方法に拡張したとも捉えられる．また，文単位の評価値を訂正単位の評価値に分配するという本研究の要求を，シャープレイ値の効率性の性質により自然に満たせることからも，シャープレイ値を用いることは十分妥当であると考える．

5.2 提案法のその他の解釈

本研究では提案法を分析的評価法と位置付けたが，提案法は以下のようにさまざまな側面への応用が考えられる基盤的な技術となりうる．

メタ評価法　提案法により得られた訂正単位の評価値と誤り訂正の性質との関係を調べることで，メタ評価に利用できる可能性がある．例えば，何らかの重要視して評価したい誤り訂正の側面があるとき，その項目に絶対値として大きな ϕ_i を与える尺度を選択することで，所望の評価を行うことができる．このことは，従来行われてきた人手評価との相関に基づいたメタ評価とは異なり，評価したいことを評価できる尺度を選択するための目的指向のメタ評価として役立つ．

バイアス・脆弱性検出法　提案法により，ある特定の訂正によって不当に性能が向上もしくは低下する問題を明らかにできる可能性がある．例えば，そもそも文法誤り訂正ではないような編集（例：文意を無視した高頻度語への置換）であっても，評価値が向上してしまうような現象が存在する可能性がある．仮にこのような現象は，悪用することで不当に高い評価値が得られる脆弱性の問題に繋がりかねない．このような脆弱性は，脆弱性と思われる訂正を含むように恣意的に作成した訂正文を用いて，提案法による分析的評価を行うことで明らかにできる可能性がある．類似した問題として，性別や国籍などに起因するバイアスを明らかにする方法にも用いることができる．これら観点においては，5.1 節で述べた議論とは異なり，人手の評価結果との比較（もしくは提案法により得られた評価値の人手評価）に意味が生じる．

説明性手法　提案法は評価値を推定した根拠を訂正の面から説明する手法であるとも見做せる．従来の説明性手法では入力の各単語に対して予測値への貢献を計算するが [15, 16]，本研究では訂正という文の部分的な変化に対して貢献を説明する点が異なるため，説明性手法としても新たな試みであると言える．

6 おわりに

シャープレイ値に基づいて，参照なし評価尺度の評価値がなぜ向上もしくは低下したのかという点を，訂正単位で分析する分析的評価手法を提案した．実際の訂正を用いた提案法の適用例を例示し，妥当性および分析的評価法以外の活用も含めた将来の展望を議論した．今後は主に 5.2 節で述べた項目について，提案法のさらなる活用について検討する．また，文法誤り訂正以外の文の編集を伴うタスク（平易化など）への活用も検討する．

参考文献

[1] Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 343–348, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[2] 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.

[3] Lloyd S Shapley, et al. A value for n-person games. 1953.

[4] 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.

[5] Md Asadul Islam and Enrico Magnani. 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**, pp. 3009–3015, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[6] 永田亮, 高村大也. 文法誤り訂正への訂正重要度の導入. 言語処理学会 第 28 回年次大会, 2022.

[7] 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.

[8] 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.

[9] Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. The CoNLL-2013 shared task on grammatical error correction. In **Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task**, pp. 1–12, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

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

[11] Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 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**, pp. 229–234, Valencia, Spain, April 2017. Association for Computational Linguistics.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

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

[14] Leshem Choshen and Omri Abend. Automatic metric validation for grammatical error correction. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1372–1382, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In **Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining**, pp. 1135–1144, 2016.

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

A シャープレイ値の具体的な計算方法

3 つの訂正 $\boldsymbol{e} = \{e_1, e_2, e_3\}$ が存在する場合を考える．．この時，各訂正の評価値 ϕ_1，ϕ_2 および ϕ_3 は次のように計算される．なお，記号および記法は 3 節と同じ意味で用いた．

$$\begin{aligned}\phi_1 =[&\frac{1}{3}(\Delta M(S, S_{\{e_1\}}) - \Delta M(S, S)) \\ &+ \frac{1}{6}(\Delta M(S, S_{\{e_1,e_2\}}) - \Delta M(S, S_{\{e_2\}})) \\ &+ \frac{1}{6}(\Delta M(S, S_{\{e_1,e_3\}}) - \Delta M(S, S_{\{e_3\}})) \\ &+ \frac{1}{3}(\Delta M(S, S_{\{e_1,e_2,e_3\}}) - \Delta M(S, S_{\{e_2,e_3\}}))]\end{aligned} \tag{3}$$

$$\begin{aligned}\phi_2 =[&\frac{1}{3}(\Delta M(S, S_{\{e_2\}}) - \Delta M(S, S)) \\ &+ \frac{1}{6}(\Delta M(S, S_{\{e_1,e_2\}}) - \Delta M(S, S_{\{e_1\}})) \\ &+ \frac{1}{6}(\Delta M(S, S_{\{e_2,e_3\}}) - \Delta M(S, S_{\{e_3\}})) \\ &+ \frac{1}{3}(\Delta M(S, S_{\{e_1,e_2,e_3\}}) - \Delta M(S, S_{\{e_1,e_3\}}))]\end{aligned} \tag{4}$$

$$\begin{aligned}\phi_3 =[&\frac{1}{3}(\Delta M(S, S_{\{e_3\}}) - \Delta M(S, S)) \\ &+ \frac{1}{6}(\Delta M(S, S_{\{e_1,e_3\}}) - \Delta M(S, S_{\{e_1\}})) \\ &+ \frac{1}{6}(\Delta M(S, S_{\{e_2,e_3\}}) - \Delta M(S, S_{\{e_2\}})) \\ &+ \frac{1}{3}(\Delta M(S, S_{\{e_1,e_2,e_3\}}) - \Delta M(S, S_{\{e_1,e_2\}}))]\end{aligned} \tag{5}$$