文法誤り訂正における
編集ベクトルの最適輸送に基づく性能評価尺度

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

文法誤り訂正の編集レベルの評価尺度では仮説編集と参照編集との一致を確認するが，表層の厳密な一致に依存しており，編集が多様化したドメインには適用が難しい．本研究では，編集をベクトル化した**編集ベクトル**を提案し，ベクトルの類似度に基づいて仮説編集と参照編集とのソフトアライメントを考える尺度 **UOT-ERRANT** を提案する．編集ベクトルは編集を適用する前後での文表現の差分ベクトルとして定義され，編集同士の類似度には最適輸送を用いている．SEEDA データセットを用いたメタ評価では従来尺度を上回る相関を示し，輸送計画が解釈性に寄与することを報告する．

1　はじめに

文法誤り訂正の参照あり評価では，図 1 に示すように，システム出力である仮説文と人手の訂正結果である参照文を比較する．この比較は誤り文からの変化を抽出した編集の単位を用いることが代表的であり，より多くの編集が仮説と参照の間で一致することが望ましい．この評価方法は ERRANT [1, 2] や PT-ERRANT [3] といった既存尺度で採用されている．しかし，いずれも編集の一致を表層の完全一致で判定していることがしばしば問題となる．図 1 では，仮説編集 [“*to*” → “*based on*”] と参照編集 [“*to*” → “*given*”] はニュアンスが一致するため仮説は完全な誤りではないと考えられるが，表層が異なるため誤りと評価されてしまう．他にも，削除するスパンの微妙な違いによっても同様の問題が生じる．

本研究では編集ベクトルを提案し，意味的な類似度を考慮しながら編集の一致を評価することでこの問題を解決する．編集ベクトルは編集前後における文埋め込みの差分として定義され，テキストに生じた変化をベクトル化する概念である．次に，仮説編集と参照編集の要素をそれぞれベクトル化し，不均衡最適輸送（Unbalanced Optimal Transport, UOT）[4] を用いて仮説から参照に編集ベクトルを輸送する評価尺度 UOT-ERRANT を提案する．計算される輸送計画は編集間のソフトアライメントとして解釈することができ，単純な正解・不正解の二値評価ではなく，部分点を適宜与えるような評価が可能である．最後に，輸送できた量とできなかった量の両方に注目することで適合率・再現率・F_β を計算し，スコアの解釈性を高める．

実験では SEEDA [5] を用いたメタ評価を実施し，UOT-ERRANT は編集レベルの尺度の中では最も高い人手評価との相関を示すこと，および輸送計画の可視化がスコアの解釈性に寄与することを示す．さらに，編集を分類した体系である誤りタイプ [2] に基づいて編集ベクトルの性質を分析したところ，類似した編集は類似したベクトルとなる結果や，ベクトルのノルムが正書法の訂正では小さく接続詞の訂正では大きいというような結果を得た．また，これらの結果は UOT-ERRANT が良い尺度となることの理由として解釈できることを議論する．

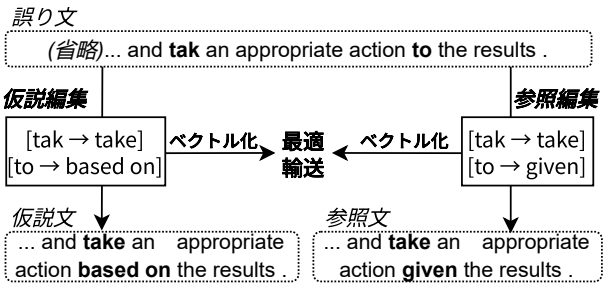

図 1: 提案する尺度 UOT-ERRANT の概念図．仮説編集と参照編集をそれぞれ抽出したのち，本稿で提案する編集ベクトルに変換し，仮説編集から参照編集にベクトルを最適輸送する．実際には，得られた輸送行列からさらに F 値を計算するが，この図では割愛した．

2 提案法

2.1 編集ベクトル

編集ベクトルを編集を適用する前後の文表現の差分として定義する．形式的には，ERRANT などの編集抽出ツールを用いて，誤り文 S とその訂正文 H の間に生じた編集を編集セット $\mathscr{E} = \{e_1, e_2, \dots\}$ として抽出する．

次に，それぞれの編集 $e \in \mathscr{E}$ を，削除することの影響を観察する leave-one-out [6] でベクトル化する．

$$V(e, \mathscr{E}, S) = \mathrm{Enc}(S_{\mathscr{E}}) - \mathrm{Enc}(S_{\mathscr{E}\setminus\{e\}}), \tag{1}$$

ここで，$S_{\mathscr{E}}$ は編集セット $\mathscr{E}$ を S に適用することを示し，$\mathrm{Enc}(\cdot) \in \mathbb{R}^d$ は d 次元の文埋め込みを計算するエンコーダである．$S_{\mathscr{E}} = H$ であるため，式 1 は訂正文から編集をただ一つ除くことによる文埋め込みの変化を，差分ベクトルとして定量化していると見做せる．単語や文の埋め込みと比較して，テキストに生じた変化（編集）をベクトル化する点が異なる．

2.2 編集のソフトアライメント

編集ベクトルに基づいて，仮説と参照との類似度を定量化する尺度 UOT-ERRANT を提案する．新たに参照文を R とし，2.1 節の手順と同様，編集抽出ツールを用いて S と H から仮説編集セット $\mathscr{E}_{\mathrm{hyp}}$ を抽出し，同時に S と R から参照編集セット $\mathscr{E}_{\mathrm{ref}}$ を抽出する．ここでの目的は，$\mathscr{E}_{\mathrm{hyp}}$ と $\mathscr{E}_{\mathrm{hyp}}$ の要素間での類似度（=ソフトアライメント）を得ることである．

この目的のために，まず，式 1 に従ってそれぞれの編集をベクトル化する．

$$\mathbf{V}^{\mathrm{hyp}} = \{V(e, \mathscr{E}_{\mathrm{hyp}}, S) | e \in \mathscr{E}_{\mathrm{hyp}}\}, \tag{2}$$

$$\mathbf{V}^{\mathrm{ref}} = \{V(e, \mathscr{E}_{\mathrm{ref}}, S) | e \in \mathscr{E}_{\mathrm{ref}}\} \tag{3}$$

次に，UOT によって仮説から参照に編集ベクトルを輸送する．UOT は 2 つの離散分布を入力として片方の分布をもう片方の分布に輸送することで，サンプル間のアライメントを定量化する．最適輸送にはいくつかの方法があるが，文法誤り訂正で頻発する過剰編集と過小編集の両方を扱うために，過不足を許容した輸送が可能な UOT を用いることとする．仮説編集セットと参照編集セットにそれぞれ n 個と m 個の編集が含まれる時，UOT の入力は，仮説編集が持つ質量 $\boldsymbol{a} = \{a_1, a_2, \dots, a_n\}$ と参照編集が持つ質量 $\boldsymbol{b} = \{b_1, b_2, \dots, b_m\}$，および編集間の輸送の難しさを示すコスト行列 $\mathbf{C} \in \mathbb{R}^{n \times m}$ の 3 点である．本稿では $\boldsymbol{a}, \boldsymbol{b}$ は編集ベクトルのノルムとし，$\mathbf{C}$ は編集ベクトル間のユークリッド距離とする．UOT の出力は輸送計画 $\mathrm{UOT}(\boldsymbol{a}, \boldsymbol{b}, \mathbf{C}) \in \mathbb{R}^{n \times m}$ である：

$$\begin{aligned}\mathrm{UOT}(\boldsymbol{a}, \boldsymbol{b}, \mathbf{C}) = \underset{\mathbf{P} \in \mathbb{R}_+^{n \times m}}{\operatorname{argmin}} & \sum_{i,j} C_{ij} P_{ij} + \epsilon H(\mathbf{P}) \\ & + \lambda_1 \mathrm{KL}(\mathbf{P}\mathbb{1}_m, \boldsymbol{a}) + \lambda_2 \mathrm{KL}(\mathbf{P}^\intercal \mathbb{1}_n, \boldsymbol{b}).\end{aligned} \tag{4}$$

ここで $\epsilon H(\mathbf{P})$ はエントロピー正則化項であり，ϵ を大きくするほど輸送が均一になる．また，λ_1, λ_2 は全ての量を輸送する・されることを強制する度合いを示す係数である．以降，計算された輸送計画 $\mathrm{UOT}(\boldsymbol{a}, \boldsymbol{b}, \mathbf{C})$ を簡単のため $\mathbf{T}$ と表記する．$T_{i,j}$ が i 番目の仮説編集から j 番目の参照編集に輸送された量を示しており，この輸送量そのものが編集のソフトアライメントを示す．

2.3 スコア計算

輸送計画 $\mathbf{T}$ から，文法誤り訂正でよく用いられる適合率・再現率・F_β を計算する．UOT においてより多くの輸送が行われるほど仮説と参照の編集が類似すると考えて，$\boldsymbol{a}, \boldsymbol{b}, \mathbf{T}$ から True Positive（TP），False Positive（FP），False Negative（FN）を定量化する．TP は仮説編集が正解である度合いであり，仮説と参照の間で輸送された総量として計算する：

$$\mathrm{Score}_{\mathrm{TP}} = \sum\nolimits_{i,j} T_{ij}. \tag{5}$$

FP は仮説編集が誤りである度合いであり，仮説から輸送されなかった量として計算する：

$$\mathrm{Score}_{\mathrm{FP}} = \sum\nolimits_{i} a_i - \mathrm{Score}_{\mathrm{TP}}. \tag{6}$$

FN は仮説編集が誤りを見逃した度合いであり，参照に輸送されなかった量として計算する：

$$\mathrm{Score}_{\mathrm{FN}} = \sum\nolimits_{j} b_j - \mathrm{Score}_{\mathrm{TP}}. \tag{7}$$

これらの値から，適合率 $\mathrm{Prec.} = \frac{\mathrm{Score}_{\mathrm{TP}}}{\mathrm{Score}_{\mathrm{TP}} + \mathrm{Score}_{\mathrm{FP}}}$，再現率 $\mathrm{Rec.} = \frac{\mathrm{Score}_{\mathrm{TP}}}{\mathrm{Score}_{\mathrm{TP}} + \mathrm{Score}_{\mathrm{FN}}}$，$F_\beta = \frac{(1+\beta^2)\mathrm{Prec.} \times \mathrm{Rec.}}{\beta^2 \mathrm{Prec.} + \mathrm{Rec.}}$ を計算する．本稿では $F_{0.5}$ を最終的なスコアとする．

3 実験

3.1 実験設定

メタ評価データセット　メタ評価[1)]のデータセットとして SEEDA [5] を用いる．SEEDA は 2 つの項目において実験設定の選択の余地があり，具体的に

1) 評価尺度を評価することをメタ評価と呼ぶ．

表 1: SEEDA-E におけるシステムレベルのメタ評価結果．r はピアソン相関，ρ はスピアマン順位相関である **太字**は各列における最大値を，下線は 2 番目に高い値を示す．

	↑ SEEDA-E Base				↑ SEEDA-E +Fluency			
	Official		10 Refs		E-Fluency		NE-Fluency	
Metrics	r	ρ	r	ρ	r	ρ	r	ρ
n-gram レベル尺度								
GLEU	.909	.965	.949	.958	.278	.600	**.781**	**.921**
GREEN	.912	.965	.910	.979	**.547**	**.802**	.745	.908
編集レベル尺度								
ERRANT	.881	.895	.952	.951	-.005	.424	.114	.508
PT-ERRANT	.924	.951	**.957**	.986	.005	.310	.276	.578
CLEME	.910	.930	.932	.937	-.043	.297	.473	.653
UOT-ERRANT	**.950**	**.979**	.942	**.993**	.445	.684	.705	.851

は (i) 人手評価の粒度：文レベル評価の SEEDA-S か編集レベル評価の SEEDA-E，(ii) ドメイン：Base 設定か+Fluency 設定，の項目である．(i) については，Kobayashi ら [5] が指摘した評価の粒度を人手評価と自動評価で一致させることの重要性に基づき，提案尺度が編集レベルの尺度であることから SEEDA-E を用いる．(ii) については両方のドメインを用いる．ドメインは，訂正システムとして GPT-3.5 [7] の出力と流暢な参照文 [8] を含むかどうかに応じて異なり，+Fluency 設定の方がより積極的に訂正するシステムを含めて評価する必要がある．

参照文　Kobayashi ら [5] が実施したように，複数の参照セットを用いて UOT-ERRANT をメタ評価する．SEEDA の Base 設定では，CoNLL-2014 共通タスク [9] の公式参照（**Official**）と，Bryant [10] らが公開する 10 種類の追加参照セット（**10 Refs**）をそれぞれ独立に用いる．+Fluency 設定では，Sakaguchi ら [8] が公開する **E-flunecy** と **NE-Flunecy** の参照セットを用いる．ただし，**E-Flunecy** に含まれる参照の一つは SEEDA のシステムセットとして使用されているため除いてから使用する．

UOT-ERRANT の実験設定．　式 1 の文表現を計算するモデル Enc(·) に ELECTRA-large [11] を用いた．この理由は，原文から置き換えられた単語を識別するという事前学習タスクが誤り検出に類似しており，誤り訂正の評価に有用な表現を計算できることを期待するからである．式 4 の $\epsilon, \lambda_1, \lambda_2$ については，Arase ら [12] に発想を得て ϵ は 0.1 に固定した．λ_1, λ_2 は GJG15 [13] と呼ばれる別のメタ評価データセットを開発データとして決定し，共に 0.1 とした．実装には，UOT 部分には POT ライブラリ [14, 15] [2)]，メタ評価部分には GEC-METRICS ライブラリ [16] を用いる．システムレベルのスコアの計算方法は，SEEDA が TrueSkill [17] を採用していることから UOT-ERRANT でも同じく TrueSkill を用いる．これは Goto ら [18] が指摘した人手評価と自動評価で計算方法を一致させることの重要性に基づく．

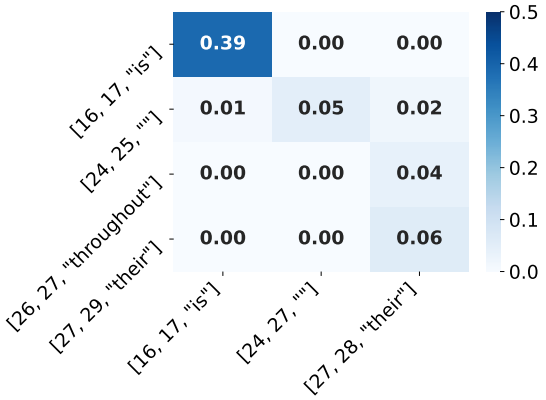

図 2: 実例における仮説編集（縦軸）と参照編集（横軸）とのソフトアライメント．実際の誤り文，参照文，仮説文は次のとおりである．
誤り文: “It is still early for parents to decide whether they can foster a new life that are not able to work and may suffer the pain in the entire life .”
参照文: “（省略）. . . new life that is not able to work and may suffer their entire life .”
仮説文: “（省略）. . . new life that is not able to work and may suffer pain throughout their life .”

比較対象の評価尺度．　既存の参照あり評価尺度と比較するために，編集レベルの既存尺度である [1, 2]，PT-ERRANT [3]，CLEME [19]，n グラムレベルの既存尺度である GLEU [20, 21]，GREEN [22] を用いる．各尺度の詳細な実験設定は付録 A にある．

3.2　実験結果

SEEDA-E におけるメタ評価結果を表 1 に示す．提案法の UOT-ERRANT は，SEEDA-E の Base 設定では他の尺度を上回る相関を達成した．+Fluency 設定では GLEU や GREEN といった n-gram レベルの尺度には劣るものの，編集レベル尺度の中では最も高い相関を達成した．編集レベル尺度はその他の粒度の尺

2) https://github.com/PythonOT/POT

表 2: 編集ベクトルの平均ノルムが大きい誤りタイプの上位 3 件（左列）と下位 3 件（右列）．ノルムは平均 $\pm$ 標準偏差 として示した．

誤りタイプ	ノルム	誤りタイプ	ノルム
正書法	0.170 $\pm$ 0.406	接続詞	1.247 $\pm$ 0.556
句読点	0.835 $\pm$ 0.597	その他	1.318 $\pm$ 0.685
副詞	0.944 $\pm$ 0.485	綴り	1.409 $\pm$ 0.610

度と比較して編集単位での良し悪しを評価できるため解釈性が高く，誤りタイプに基づいて評価できるなどの説明性に関する利点がある．UOT-ERRANT はこの説明性の利点を維持しながら，特に+Fluency 設定において相関を向上させていることがわかる．

また，輸送計画 $\mathbf{T}$ の可視化はスコアの解釈に有用である．図 2 は輸送計画の実例であり，縦軸に示される仮説編集と，横軸に示される参照編集とのアライメントを示す．各編集は，誤り文への単語レベルスパンとその訂正後の文字列として表記した．従来の評価は誤り文へのスパンと訂正後の文字列の完全一致に基づくため，[16, 17, “*is*”] の仮説編集しか正しいと評価できなかった．一方，提案法では仮説編集 [24, 25, “”] と参照編集 [24, 27, “”]，もしくは仮説編集 [27, 29, “*their*”] と参照編集 [28, 29, “*their*”] のように訂正後の文字列は一致するがスパンが僅かに異なる編集にも評価値を与える．このように部分点を与えられる仕組みが，人間の直感とより一致する評価に繋がったと考えられる．

4 分析：編集ベクトルの性質

編集ベクトルの性質を ERRANT が定義する誤りタイプ [2] に従い分析する．誤りタイプは名詞の数や冠詞，動詞の時制などの編集の分類を示すものである．まず，誤りタイプごとにノルムの平均を計算した．表 2 にノルムの大きさが上位と下位の 3 件である誤りタイプを示した．正書法（空白や文字の大小）や句読点に関する編集のノルムは小さく，接続詞や綴りに関する編集のノルムは大きい結果であった．この結果から，ノルムは編集が持つ意味変化の度合いを暗黙的に示していると言える [3]．UOT-ERRANT では，各編集の輸送量 $\boldsymbol{a}, \boldsymbol{b}$ にこのノルムを用いることで編集を暗黙的に重み付けていると解釈できる．特に，表 2 に見られる正書法を軽視する傾向はニューラルモデルに基づく参照なし尺度

3) 綴りは誤り方によっては文意の把握に影響すると考えられる．

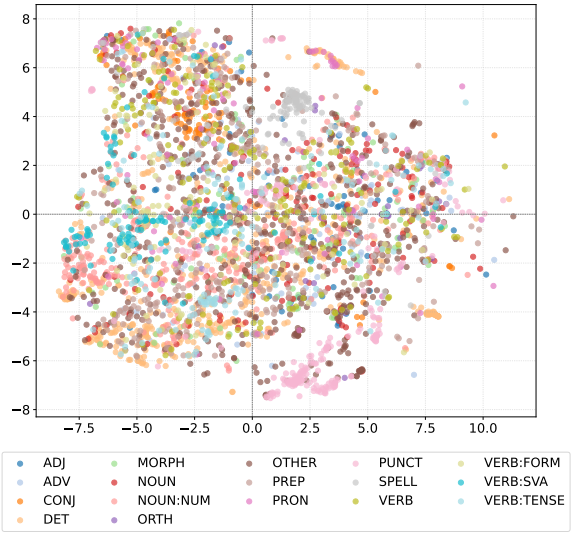

図 3: 編集ベクトルを t-SNE で次元削減し，誤りタイプごとに色分けした散布図．

にも見られており [23]，人手評価と高く相関する評価に貢献するものと考える．

次に，類似の編集は類似の編集ベクトルになるかを確かめるため，誤りタイプごとに編集ベクトルがクラスタを形成するかを調べた．図 3 は，t-SNE [24] を用いて編集ベクトルを次元削減し，誤りタイプで色分けした散布図である．なお，SEEDA の GPT-3.5 の出力から抽出した編集を用いた．この結果から，句読点（PUNCT），綴り（SPELL），主語動詞の一致（VERB:SVA）はクラスタを形成した．一方，NOUN や VERB などの内容語に関する編集は，語彙によって意味変化の度合いが異なるため，編集ベクトルは分散する傾向にあった．この性質は，UOT におけるコスト行列 $\mathbf{C}$ が正確に編集同士の近さを定量化することに貢献すると考えられる．

5 おわりに

テキストの変化をベクトル化する編集ベクトルを提案し，仮説編集から参照編集に編集ベクトルを最適輸送するアイデアに基づいた評価尺度 UOT-ERRANT を提案した．編集ベクトルは編集間の類似度計算を可能にし，本研究で実施したような評価尺度に加えて，編集の用例検索など幅広い応用が期待される手法である．この特徴は，編集をスカラとして定量化してきた既存手法 [25, 19, 26] には見られない．今後は編集ベクトルのさらなる応用や，UOT-ERRANT を活用した GEC システムの評価と分析に取り組む．

謝辞

本研究は JST 次世代研究者挑戦的研究プログラム JPMJSP2140 の支援を受けたものです.

参考文献

[1] Mariano Felice, Christopher Bryant, and Ted Briscoe. Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In Yuji Matsumoto and Rashmi Prasad, editors, **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.

[2] Christopher Bryant, Mariano Felice, and Ted Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In Regina Barzilay and Min-Yen Kan, editors, **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.

[3] 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.

[4] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 28. Curran Associates, Inc., 2015.

[5] 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, 2024.

[6] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 3719–3728, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[8] Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 169–182, 2016.

[9] 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 Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant, editors, **Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task**, pp. 1–14, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[10] Christopher Bryant and Hwee Tou Ng. How far are we from fully automatic high quality grammatical error correction? 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 1: Long Papers)**, pp. 697–707, Beijing, China, July 2015. Association for Computational Linguistics.

[11] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[12] Yuki Arase, Han Bao, and Sho Yokoi. Unbalanced optimal transport for unbalanced word alignment. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3966–3986, Toronto, Canada, July 2023. Association for Computational Linguistics.

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

[14] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. **Journal of Machine Learning Research**, Vol. 22, No. 78, pp. 1–8, 2021.

[15] Rémi Flamary, Cédric Vincent-Cuaz, Nicolas Courty, Alexandre Gramfort, Oleksii Kachaiev, Huy Quang Tran, Laurène David, Clément Bonet, Nathan Cassereau, Théo Gnassounou, Eloi Tanguy, Julie Delon, Antoine Collas, Sonia Mazelet, Laetitia Chapel, Tanguy Kerdoncuff, Xizheng Yu, Matthew Feickert, Paul Krzakala, Tianlin Liu, and Eduardo Fernandes Montesuma. Pot python optimal transport (version 0.9.5), 2024.

[16] Takumi Goto, Yusuke Sakai, and Taro Watanabe. gec-metrics: A unified library for grammatical error correction evaluation. In Pushkar Mishra, Smaranda Muresan, and Tao Yu, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)**, pp. 524–534, Vienna, Austria, July 2025. Association for Computational Linguistics.

[17] 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.

[18] Takumi Goto, Yusuke Sakai, and Taro Watanabe. Rethinking evaluation metrics for grammatical error correction: Why use a different evaluation process than human? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 1165–1172, Vienna, Austria, July 2025. Association for Computational Linguistics.

[19] Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. CLEME: Debiasing multi-reference evaluation for grammatical error correction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 6174–6189, Singapore, December 2023. Association for Computational Linguistics.

[20] 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.

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

[22] 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.

[23] Takumi Goto, Justin Vasselli, and Taro Watanabe. Improving explainability of sentence-level metrics via edit-level attribution for grammatical error correction. In Jin Zhao, Mingyang Wang, and Zhu Liu, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)**, pp. 1004–1015, Vienna, Austria, July 2025. Association for Computational Linguistics.

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. **Journal of Machine Learning Research**, Vol. 9, No. 86, pp. 2579–2605, 2008.

[25] 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.

[26] Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. IMPARA: Impact-based metric for GEC using parallel data. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 3578–3588, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

A　実験設定の詳細

3.1 節で UOT-ERRANT と比較した尺度についての実験設定を下記に示す．複数参照の評価設定においては，特に記載のない限り，参照が複数ある場合は最も高いスコアが得られる参照を文ごとに選択することができる．

ERRANT.適合率を重視する $F_{0.5}$ を用いた．

PT-ERRRANT.ERRANT の計算において，編集を BERTScore で重み付ける評価尺度である．BERTScore の計算には `bert-base-unrcased` を用いて，F_1 スコアを計算した．ベースラインによる rescaling は実施し，IDF による追加の重み付けは行わない．編集の抽出は ERRANT と同様の実装を用いた．最終的なスコアは重みづけられた $F_{0.5}$ である．

CLEME.CLEME-indenpent の方法を用いて，True Positive，False Positive，False Negative，True Negative それぞれに対する scale factor と threshold は全て原論文の設定に従った．

GLEU.n を $n = \{1, 2, 3, 4\}$ として計算した．参照が複数ある場合は各参照におけるスコアの平均を用いた．

GREEN.GLEU と同様に，n を $n = \{1, 2, 3, 4\}$ として計算した．Koyama ら [22] の報告に従い $F_{2.0}$ を用いた．