IMPARA: パラレルデータにおける修正の影響度に基づいた文法誤り訂正の自動評価法

前田 航希　金子 正弘　岡崎 直観
東京工業大学
{koki.maeda@nlp., masahiro.kaneko@nlp., okazaki@}c.titech. ac.jp

概要

文法誤り訂正（Grammatical Error Correction; GEC）の自動評価は，低コストかつ定量的な評価に不可欠である．しかし，既存の GEC 自動評価手法は評価時に複数の参照文を必要としたり，評価モデルの学習に特化した訓練データが必要になるなど，自動評価の実現のためのデータ作成コストが高いという難点がある．本稿では，誤文と正文の組からなるパラレルデータのみを用い，修正の影響度を考慮しながら GEC の評価尺度を学習する手法である IMPARA を提案する．提案手法は GEC の自動評価におけるデータ作成コストを大幅に軽減しつつ，人手評価との相関において既存手法と同等以上の性能を示した．また，評価尺度を学習するパラレルデータを変更することで，異なるドメインや訂正スタイルに適合した評価を実現できることを実験的に示した．

1　はじめに

GEC は，文法的な誤りを含んだ文（誤文）が入力されたとき，文法的に正しい文（正文）となるように訂正して出力するタスクである．GEC はウェブテキスト [1] や言語学習者が書いたエッセイ [2] など，様々なドメインで活用される．GEC の精度は，システムが出力した文が文法的に正しいと人間に受け入れられる度合いで測定できる．ただ，人手評価はコストが高いため，人手評価と相関の高い自動評価手法の確立が望まれている．

GEC の自動評価手法は二つに大別される．一つは，誤文を含むコーパスに対して人間が正文を付与し，GEC システムの出力との近さを評価する参照文有りの手法 [3, 4, 5] である．一般に，誤文の訂正の仕方は複数あり得るため，参照文有りの手法で正しい評価を行うには，複数の参照文を用意する必要がある．ところが，考えうるすべての訂正を網羅するのは困難である．とはいえ，参照文の数を絞ってしまうと，GEC システムの訂正が正しいとしても，参照文と訂正結果に差異があれば過小評価される．

もう一つは，入力文とシステム出力のみを用いる参照文なしの評価手法である．参照文なしの評価手法として，言語モデルに基づく手法 [6, 7, 1] があるが，GEC に関するデータを全く用いないため，人間による評価との相関が低いことが知られている [8]．これに対し，GEC システムの出力文を人手で評価した結果を活用し，評価尺度を最適化する手法 SOME が提案されている [8]．しかし，SOME の訓練データを作成するには，GEC システムの出力文に対して人手評価を付与する必要がある．そのため，異なるドメインや訂正スタイルに対応するには，新たに人手評価データを作成しなければならない．

本研究では，評価時に参照文を用いず，誤文と正文から構成されるパラレルデータのみから GEC の自動評価尺度を学習できる **IMPARA**（IMpact-based metric for GEC using PARAllel data）を提案する．評価尺度の学習では，GEC システムの訓練データと同じ形式のパラレルデータを利用できるため，データ作成コストを大幅に削減できる．また，自動評価尺度の学習に用いるパラレルデータを変えることで，様々なドメインや訂正スタイルの特徴を考慮した自動評価モデルを構築できる．

メタ評価実験では，(i) IMPARA が人手評価データを用いた既存の評価手法と比較して同等または上回る評価性能を持つこと，(ii) 異なるドメインや訂正スタイルのデータで評価尺度を学習することで，ドメインやスタイルの特性を考慮した自動評価手法を構築できることを実証する．

2　提案手法（IMPARA）

IMPARA は，図 1 のように出力文を評価する訂正評価モデルと，入力文と出力文の類似性を評価する

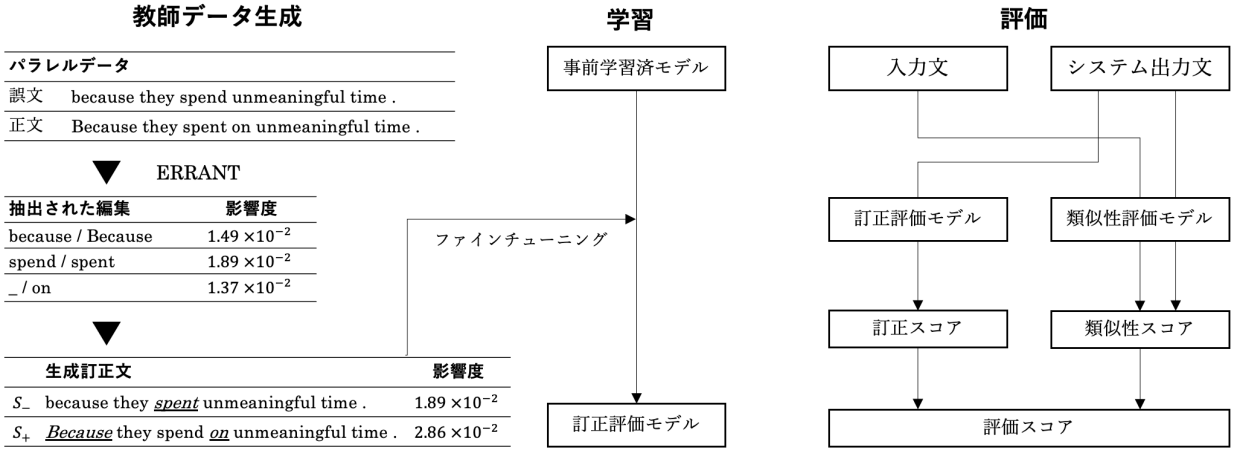

図 1　IMPARA による評価スコアの自動算出（右側）と評価尺度の学習方法（中央）および訓練データの自動作成（左側）

類似性評価モデルから構成される．

訂正評価モデルは訂正後の文に対する相対的な評価を付与する．誤文に対する編集操作の適用数により文対を順位付けし，GEC 評価指標をメタ評価する既存研究 [9] に着想を得て，誤文に対して編集操作をランダムに適用した文対を比較し，優劣の順序関係を学習する．誤文から正文への書き換えでは，編集操作ごとに異なる深刻さの誤りを訂正していると考え，編集操作の**影響度**を定義し，編集操作を適用した文対に関する順序関係を決定する．

類似性評価モデルは入力文と出力文の内容が乖離することを防止する．既存の参照文なし評価手法 [10] では，入力文と出力文の表層的な一致率を計算していたが，IMPARA では文ベクトルの類似度を計算する．これにより，文意を変更しない表層的な訂正に対しても，類似性を頑健に判定できる．

2.1　編集の影響度

編集操作に対して影響度を形式的に定義する．誤文 S と正文 T の組を (S,T) とする．$\mathscr{E}$ を S から T への順序によらない編集操作の集合とする．編集を適用する関数 f とし，S に対して $\mathscr{E}$ に含まれる全ての編集を適用して T を得ることを，$T = f(S,\mathscr{E})$ と書く．ある編集操作 $e \in \mathscr{E}$ が文の意味を大きく変化させるならば，その編集操作の影響度は高いと考える．そこで，編集 e を除外した文 $T_{-e} = f(S,\mathscr{E}\setminus\{e\})$ と正文 T との距離を考え，e の影響度 I_e を定義する．

$$I_e = 1 - \frac{\mathrm{BERT}(T)\cdot\mathrm{BERT}(T_{-e})}{\|\mathrm{BERT}(T)\|\|\mathrm{BERT}(T_{-e})\|} \tag{1}$$

ここで，$\mathrm{BERT}(T)$ は文 T に対して事前学習済み BERT が計算する文ベクトルであり，文 T の全トークンの最終層のベクトルの平均である．また編集の部分集合 $E \subseteq \mathscr{E}$ を適用した文 $f(S,E)$ の影響度を E に含まれる各編集の影響度の総和 $\sum_{e\in E} I_e$ とする．

2.2　自動評価手法の構築

IMPARA は入力文 S と出力文 O に対して訂正評価モデルと類似性評価モデルの両方のスコアを考慮し，評価スコア $\mathrm{score}(S,O) \in [0,1]$ を計算する．

$$\mathrm{score}(S,O) = \begin{cases} \mathrm{corr}(O)\ (\text{if } \mathrm{sim}(S,O) > \theta) \\ 0\ (\text{otherwise}) \end{cases} \tag{2}$$

ただし，$\mathrm{sim}(S,O)$ は類似性スコア，$\mathrm{corr}(O)$ は訂正スコア，θ は類似性スコアの閾値である．類似性スコアが閾値 θ 以下である場合は，出力文が入力文に対する適切な訂正ではないと判断し，評価スコアを 0 とする．一方，類似性スコアが閾値よりも大きければ，訂正スコアの値をそのまま採用する．

訂正スコア　訂正スコアは，出力文 O を訂正評価モデル R に与え，シグモイド関数 σ を用いて $\mathrm{corr}(O) = \sigma(R(O))$ で求める．訂正評価モデル R は BERT の先頭トークンに線形変換を適用し，実数値を計算するもので，ファインチューニングにより構築する．ファインチューニングのための訓練データは，誤文と正文のパラレルデータから自動的に作成する．n 個の誤文と正文の対からなるパラレルデータを $\mathscr{C} = \{(S_i,T_i)\}_{i=1}^{n}$ と表現する．誤文と正文の組 $(S,T) \in \mathscr{C}$ に対し，ERRANT [5] を用いてアライメントを計算し，編集集合 E を抽出する．E から異なる編集の部分集合 E',E'' を作成する．E',E'' を S に適用することで新たな文対を生成し，式 1 の影響度でその順序関係を決定し，R の訓練データとする．

訓練データの作成手順を説明する．誤文と正文の組 (S,T) に対して抽出された編集集合を

$E = \{e_1, \ldots, e_{|E|}\}$ とし，$k \in \{1, 2, \ldots, |E|\}$ を離散一様分布から選ぶ．編集操作 $e \in E$ を k 回非復元抽出し，抽出された編集操作の集合を E' とする．続いて，E' を微修正することで E'' を作成する．まず $E'' = E'$ と初期化する．その後，E の各要素 $e \in E$ について，確率 $\frac{1}{|E|}$ で以下の操作を適用する．

$$E'' \leftarrow \begin{cases} E'' \cup \{e\} \text{ if } e \notin E' \\ E'' \setminus \{e\} \text{ if } e \in E' \end{cases} \tag{3}$$

E', E'' を誤文に適用して得た文 $f(S, E'), f(S, E'')$ に対して影響度を計算し，影響度の低い文を S_-，高い文を S_+ とする．このように (S, T) から文対 (S_-, S_+) を作成し，訂正評価モデル R の訓練データ $\mathcal{T}$ を得る．ただし，一つの文対から作成する訓練事例は最大で c 件とする．また，E' と E'' が等しくなった場合は訓練事例として採用しない．

得られた訓練データ $\mathcal{T}$ を用いて訂正評価モデル R を学習する．訂正文の順序関係を学習するため，モデル R の損失関数 L を次式で定義する．

$$L = \frac{1}{|\mathcal{T}|} \sum_{(S_-, S_+) \in \mathcal{T}} \sigma\left(R(S_-) - R(S_+)\right) \tag{4}$$

損失関数にシグモイド関数 σ を用いるのは一部の事例を極端に重要視することを防ぐためで，実験的に評価性能の向上に寄与することを確認している．

類似性スコア　入力文 S と出力文 O をそれぞれ事前学習された BERT に与え，類似性スコア $\mathrm{sim}(S, O)$ を計算する．影響度と同様に，文中の全てのトークンの最終層のベクトルの平均として文ベクトルを算出し，両文ベクトルのコサイン類似度を $\mathrm{sim}(S, O)$ とする．これにより，入力文 S と出力文 O の情報がどの程度類似しているかを計測できる．

3　実験

IMPARA の評価スコアと人間の評価結果との相関を調べるため，二つのメタ評価実験を行った．まず，GEC システムに対する人手評価と，自動評価手法の評価スコアとのピアソンの積率相関係数（Pea）およびスピアマンの順位相関係数（Spe）を計測する．また，文の優劣比較の正解率（Acc）とケンドールの順位相関係数（Ken）を計測する．人手評価として，CoNLL2014 [11] データセットに対し複数の GEC システムの訂正文を人間に順位付けさせたデータセット[1)]を用いる．

次に，ドメインや訂正スタイルを考慮した評価スコアを構築できたか検証する．この検証では，IMPARA の訂正評価モデルの学習に用いるコーパスと，評価尺度のメタ評価に用いるコーパスの組合せを変え，それぞれについて MAEGE [9] によるメタ評価を実施する．MAEGE は，評価データの誤文に加えた編集数に基づく順序関係を構築する．この順序関係を用いて作成した評価スコアと自動評価手法の評価スコアの相関係数を測定してメタ評価を行う．この実験には公開されている実装[2)]を利用した．

3.1　実験設定

人手評価との相関を測定する実験では，IMPARA の訂正評価モデルの学習に CoNLL2013 [13] を用いた．ドメインや訂正スタイルを変更する実験では，CoNLL2014 [11] に加え，ウェブテキストの文法誤りを訂正する CWEB [1]，流暢な訂正を含む JFLEG [14]，エッセイに訂正を加えた FCE [2] を用いた．いずれのデータセットも 90%をモデルの学習に利用し，10%をメタ評価のためのデータとした．

実験では，Hugging Face が公開する事前学習済モデル[3)]（BERT-BASE-CASED）を BERT モデルとして採用した．類似性評価モデルは事前学習済モデルをそのまま用い，訂正評価モデルは自動生成した訓練データでファインチューニングした．

ベースライン手法として参照文を用いない自動評価手法である SOME と Scribendi Score [10] を用いた．また，IMPARA の訂正評価モデルの訓練データ構築法の有効性を検証するため，パラレルコーパスの誤文と正文の文対のみを用いて BERT をファインチューニングした評価モデル（パラレルのみ）と比較した．SOME の評価モデルとして用いられる BERT のファインチューニングでは，公開されたデータセット[4)]を本研究と同じ学習・評価の分割で利用し，吉村ら [8] のハイパーパラメータ設定を利用した．

3.2　実験結果

各自動評価と人手評価との相関を測定した結果を表 1 に示す[5)]．IMPARA はコーパス単位の評価では

1）　本実験では，Grundkiewicz ら [12] の論文の Table 3(b) に示される Expected Wins を人手評価とした．

2）　https://github.com/borgr/EoE
3）　https://github.com/huggingface/transformers
4）　https://huggingface.co/datasets/tmu_gfm_dataset
5）　Scribendi Score は Islam・Magnani [10] の論文で報告されている評価値を再現できなかったため，論文の Table 3 の値を併記する．

表 1 Grundkiewicz の人手評価との相関

	コーパス		文	
	Pea	Spe	Acc	Ken
IMPARA	**0.974**	**0.934**	0.748	0.496
パラレルのみ	0.936	0.929	0.742	0.485
SOME	0.956	0.923	**0.777**	**0.555**
Scribendi	0.303	0.729	0.414	-0.170
Scribendi 論文	0.951	0.940	—	—

人手評価データを用いる SOME より高い相関を示した．さらに，パラレルコーパスに含まれる文対のみを訂正評価モデルの訓練データとした場合との比較から，提案手法で訓練データを自動構築することの効果を確認できた．また，MAEGE によるメタ評価によると，IMPARA はコーパス単位の評価で既存手法と同等程度，文単位の評価で既存手法に対して 0.1 ポイント以上高い性能を示した（Appendix の表 4）．これらの実験結果は，IMPARA はパラレルデータのみから自動構築した訓練データのみを用いるが，人手評価を付与した訓練データを用いる既存の評価手法と同等程度の評価性能を達成することを示唆している．

4 種類の評価コーパスに対して，学習コーパスを変化させたときの MAEGE によるメタ評価を行った結果を表 2 に示す．実験の結果，評価に用いたコーパスと同じコーパスを用いて訂正評価モデルを学習させることで評価性能が向上した．特に文単位の相関が大きく向上していることが確認された．さらに，MAEGE を用いて既存手法と評価性能を比較した（Appendix の表 5）．SOME は CWEB, FCE において，Scribendi Score は CWEB, JFLEG において評価性能が極端に低下した．一方で，IMPARA はいずれの評価コーパスにおいても高い評価性能を示した．このことから，IMPARA を用いることでコーパスの特性により適合した評価を行えることが確認できた．

3.3 誤りタイプごとの影響度

2.1 節で定義した影響度を用いたとき，BERT モデルがどのような誤りに対して高い影響度を算出するのか分析した．CoNLL2014 に含まれる訂正文対について，ERRANT を用いて編集と誤りタイプを抽出し，編集ごとに影響度を求め，平均値を算出した．

平均影響度を算出した誤りタイプについて，OTHER を除き事例が 400 以上の誤りタイプを抜粋して表 3 に示す（全ての結果は Appendix の表 6）．DET（冠詞），PREP（前置詞）などの機能語と比較して，NOUN（名詞），VERB（動詞）など内容語の訂正に高い影響度を付した．また同じ名詞に関する訂正でも，数量に関する訂正には低い影響度が算出されることが確認された．以上から，本研究で定義した影響度は文法的な役割に関する訂正よりも内容語による意味の変化を重要視していると考えられる．

表 2 学習・評価コーパスの組合せによる性能変化

評価データ	学習データ	コーパス		文		束
		Pea	Spe	Pea	Spe	Ken
CoNLL2013	CoNLL2013	0.932	**1.000**	**0.411**	**0.515**	**0.688**
	CWEB	0.961	**1.000**	0.380	0.468	0.574
	JFLEG	0.959	0.990	0.344	0.408	0.568
	FCE	**0.967**	**1.000**	0.404	0.490	0.567
CWEB	CoNLL2013	0.750	0.836	0.331	0.328	0.713
	CWEB	0.790	**0.963**	**0.472**	**0.432**	**0.780**
	JFLEG	0.757	0.818	0.353	0.354	0.775
	FCE	**0.805**	0.936	0.350	0.397	0.775
JFLEG	CoNLL2013	0.959	0.990	0.516	0.604	0.677
	CWEB	0.952	0.972	0.524	0.572	0.644
	JFLEG	0.937	**1.000**	**0.618**	**0.685**	**0.783**
	FCE	**0.961**	0.990	0.581	0.649	0.627
FCE	CoNLL2013	0.865	0.972	0.377	0.388	0.758
	CWEB	**0.882**	**0.990**	0.435	0.441	0.753
	JFLEG	0.852	0.972	0.390	0.429	0.739
	FCE	0.853	**0.990**	**0.541**	**0.616**	**0.848**

表 3 CoNLL2014 において，OTHER を除く事例数が 400 以上の誤りタイプ

誤りタイプ	説明	影響度 (10^{-2})	事例数
NOUN	名詞	0.652	408
VERB:TENSE	動詞の時制・態	0.649	480
VERB	動詞	0.580	557
NOUN:NUM	名詞の数量	0.385	534
PUNCT	句読点	0.367	473
DET	冠詞	0.364	1142
PREP	前置詞	0.325	700

4 おわりに

本稿では，パラレルコーパスを用いた GEC の自動評価手法の構築法である IMPARA を提案した．提案手法は，編集に対して影響度を算出し，訂正文の良し悪しに関する順序関係を自動的にラベル付けした訓練データを生成し、BERT をファインチューニングすることで，訂正文に対する相対的な評価尺度を獲得する．提案手法は，人手評価との相関において既存手法と同等以上の性能を示すこと，評価対象となるコーパスが持つ訂正の特性を考慮した自動評価を行えることを確認した．

今後の発展として，文法誤り生成を用いた訓練データ生成を取り入れることがある．データ生成コストの緩和が見込まれるほか，自動評価尺度の品質向上に寄与すると考えられる．

参考文献

[1] Simon Flachs, Ophélie Lacroix, Helen Yannakoudakis, Marek Rei, and Anders Søgaard. Grammatical error correction in low error density domains: A new benchmark and analyses. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 8467–8478, Online, November 2020. Association for Computational Linguistics.

[2] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 180–189, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[3] Daniel Dahlmeier and Hwee Tou Ng. Better evaluation for grammatical error correction. In **Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 568–572, Montréal, Canada, June 2012. Association for Computational Linguistics.

[4] 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 (Volume 2: Short Papers)**, pp. 588–593, Beijing, China, July 2015. Association for Computational Linguistics.

[5] 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.

[6] Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2109–2115, Austin, Texas, November 2016. Association for Computational Linguistics.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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 **Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task**, pp. 1–14, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[12] 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.

[13] 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.

[14] 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.

表 4 MAEGE によるメタ評価

	コーパス		文		束
	Pea	Spe	Pea	Spe	Ken
IMPARA	0.951	0.990	**0.522**	**0.608**	0.692
SOME	**0.965**	**1.000**	0.394	0.439	0.563
Scribendi	0.884	0.981	0.374	0.421	**0.824**

表 5 評価データと同じドメインの訓練データで学習した IMPARA と既存手法の性能比較

評価データ	手法	コーパス		文		束
		Pea	Spe	Pea	Spe	Ken
CoNLL2013	IMPARA	0.932	**1.000**	**0.411**	**0.515**	0.688
	SOME	**0.961**	**1.000**	0.370	0.419	0.502
	Scribendi	0.938	0.984	0.331	0.355	**0.698**
CWEB	IMPARA	**0.790**	**0.963**	**0.472**	**0.432**	**0.780**
	SOME	0.767	0.663	0.055	0.155	0.678
	Scribendi	0.637	0.451	0.177	0.194	0.616
JFLEG	IMPARA	0.937	**1.000**	**0.618**	**0.685**	**0.783**
	SOME	**0.955**	0.990	0.523	0.531	0.639
	Scribendi	0.932	0.945	0.255	0.303	0.574
FCE	IMPARA	0.853	**0.990**	**0.541**	**0.616**	0.848
	SOME	0.843	0.972	0.165	0.254	0.663
	Scribendi	**0.869**	0.933	0.342	0.449	**0.897**

表 6 誤りタイプ毎の影響度と事例数（CoNLL2014）

誤りタイプ	説明	影響度 (10^{-2})	事例数
OTHER	非分類	1.322	1191
CONTR	縮約	1.230	23
SPELL	つづり	1.174	249
ORTH	空白や大小文字	1.011	99
NOUN	名詞	0.652	408
VERB:TENSE	動詞の時制・態	0.649	480
VERB	動詞	0.580	557
ADV	副詞	0.560	130
CONJ	接続詞	0.538	53
ADJ	形容詞	0.526	131
PRON	代名詞	0.465	227
MORPH	品詞	0.451	229
NOUN:NUM	名詞の数量	0.385	534
PUNCT	句読点	0.367	473
DET	冠詞	0.364	1142
PREP	前置詞	0.325	700
VERB:FORM	不定詞・動名詞	0.319	262
PART	前置詞	0.300	99
VERB:SVA	三人称単数現在	0.284	291
ADJ:FORM	形容詞の比較級	0.284	16
NOUN:POSS	名詞の所有格	0.269	53
NOUN:INFL	名詞の可算・不可算	0.256	14
WO	語順	0.238	53
VERB:INFL	時制・態の誤用	0.065	6

A ハイパーパラメータの設定

訂正評価モデルの再学習では，比較時にコーパスの量による影響を回避するために，コーパスによらず $|\mathcal{T}| = 4096$ となるように調整した．コーパスに含まれる訂正文対 1 つから生成する文対の最大数 c を 30，学習率を 10^{-5} とし，バッチサイズは 32 とした．エポック数は $1, \ldots, 10$ と変化させ，モデルを学習した．類似性スコアの閾値 θ は 0.9 とした．