Vol.2017-NL-234 No.3
2017/12/20

文法誤り訂正の文単位評価におけるリファレンスレス手法の評価性能

浅野 広樹[1,2,a)]　水本 智也[2,b)]　松林 優一郎[1,c)]　乾 健太郎[1,2,d)]

概要：文法誤り訂正の自動評価はリファレンスを使わないリファレンスレス評価手法が浅野らによって提案され，リファレンスレス手法の評価性能は従来のリファレンスベース手法を上回ったとされている．その評価性能は，あるテスト文書に対する複数の訂正システムの出力全体に対する自動評価スコアと人手評価スコアの相関によって評価された．しかし，文書全体に対するスコアの相関が高いからといって，文単位での評価も適切であるとは限らない．文単位評価が適切にできれば，訂正システムの誤り分析を効率的に行うことができる．そこで本研究では，文法誤り訂正の自動評価の文単位での評価性能を調査した．その結果，リファレンスレス手法の評価性能が文単位でも従来のリファレンスを用いる手法を上回った．このことは，リファレンスレス手法は複数の訂正候補からよい訂正を選べることを意味するため，訂正システムに応用できる可能性がある．実際に，複数の訂正システムの出力をリファレンスレス評価手法で評価し最もよい訂正を選択することによって訂正性能が向上することを確かめた．

キーワード：文法誤り訂正，自動評価，文単位評価

1. はじめに

文法誤り訂正（Grammatical Error Correction: GEC）は，学習者の書いた文を文法的な文に訂正するタスクである．GEC は本質的には機械翻訳や自動要約などと同様に生成タスクであり，1 つの入力に対する出力の正解が一つだけとは限らずその自動評価は難しい．そのため，GEC の自動評価は重要な課題であり自動評価尺度に関する研究が多く行われてきた．

GEC の自動評価尺度は，大きく分けて二種類提案されている．一つは正解データ (リファレンス) を使った手法である．リファレンスを使った手法であるため，本稿ではこの手法をリファレンスベース評価尺度と呼ぶ．もう一つは正解を使わずに評価する手法である．これはリファレンスを使わない評価尺度のため，リファレンスレス評価尺度と呼ぶ．リファレンスベース手法は，システムの出力とリファレンス文を比較することで評価する [5], [9], [15]．一方，リファレンスレス手法は，システムの出力のみもしくはシステムの出力と学習者の書いた文を使って評価する [13], [16]．

これらの GEC の自動評価尺度の評価性能は，自動評価尺度が算出した誤り訂正の質に関するスコアを，人手により算出したスコアと比較することで検証される．具体的には，図 1 右上のように複数の GEC システムを用いて，まず，それぞれのシステムが出力した訂正文に対して人手による評価スコアと自動評価スコアの平均を計算する．つぎに，複数のシステムについてのこれらの平均スコアを用いて，人手によるスコアのランキングと自動評価の平均スコアのランキングの順位相関係数を求めることで自動評価尺度の評価性能を検証する．この評価性能の検証手法では，システム単位でスコアの平均を取って比較されることから，本稿ではこの手法をシステム単位評価と呼ぶ．

我々は [13] において，我々の提案する新しいリファレンスレス評価尺度がシステム単位評価においてリファレンスベースの評価尺度を上回ることを報告した．しかし，これまで，GEC の評価性能の検証する研究においては，リファレンスベース，リファレンスレス評価尺度共に，文単位で見た場合に人間の評価に近い結果を自動評価で算出できるかは検証されていない．つまり，ある文に対して二つの訂正結果が与えられた場合に，自動評価尺度でより優れた訂正を高く評価できるかは明らかなっていない．

[1] 東北大学
Tohoku University
[2] 理化学研究所
RIKEN
[a)] asano@ecei.tohoku.ac.jp
[b)] tomoya.mizumoto@riken.jp
[c)] y-matsu@ecei.tohoku.ac.jp
[d)] inui@ecei.tohoku.ac.jp

文単位での評価が可能になれば，GEC システムの人手による誤り分析に有用である．GEC システムを改善する開発者に対して，手法ごとに正しく訂正できた，誤った訂正をしてしまった例を提示することができる．これにより効率的な誤り分析が可能になる．

上記までの背景をふまえ，現在提案されている自動評価尺度が文単位でどの程度頑健に評価できるかを調査する．システム単位評価に対して，文単位で評価性能を検証するため文単位評価と呼ぶ．文単位の評価値は図 1 右下のように，ある文に対して任意の訂正文ペアを抽出し，自動評価スコアが人手スコアの順位を再現できるかの正答率により算出する．文法誤り訂正の自動評価尺度に対して，この文単位での性能調査を行うのは本稿が最初である．結果として，文単位評価においても，リファレンスレス評価尺度がリファレンスベース評価尺度よりも優れていることがわかった．

正解を使わないリファレンスレス評価尺度が文単位でもリファレンスベースよりも良い評価ができる結果を受け，本稿ではリファレンスレス評価尺度のもう一つの可能性を調査するために実験を行う．リファレンスレス評価尺度は正解データを必要としないため，正解データのない文に対しても評価スコアを与えることができる．つまり，リファレンスレス評価尺度を使えば，GEC システムの出力した訂正文の候補の中から最も良い訂正文を選択することで誤り訂正ができる可能性がある．これを確かめるために，複数の GEC システムの出力に対してリファレンスレス評価尺度でスコアを付与し，最もスコアの高いシステムの訂正文を採用するアンサンブル手法で，誤り訂正の性能を調査した．実験の結果，人手の評価，M^2 および GLEU でアンサンブル手法がアンサンブルする前のシステムを上回ることがわかった．

2. 既存の評価尺度

本節では，これまで提案された GEC のシステム単位評価における自動評価尺度とその評価尺度の評価性能を検証するために行われてきた方法について説明する．

2.1 リファレンスベース手法

訂正システムの評価では，学習者の書いた文に対して人手で訂正したリファレンスを使うことが一般的である．このリファレンスベース評価は M^2 [5]，I-measure [9]，GLEU+ [15], [20] が考案されている．

2.1.1 M^2

GEC の初期の研究では，訂正システムが行った編集操作がどの程度正解の編集と一致しているかを F 値で評価していた [6], [7]．しかし，長いフレーズの編集が必要な場合などに訂正システムを過小評価してしまうという問題があった．M^2 は ”edit lattice” を用いることにより，システムが行った編集操作を正解と最大一致するように同定する手法である．M^2 によって算出された $F_{0.5}$ 値が CoNLL 2014 Shared Task on GEC で採用されて以降，文法誤り訂正の評価尺度として最も用いられている．

2.1.2 I-measure

M^2 の問題点の一つに，訂正を全く行わないシステムと誤った訂正のみを出力するシステムに対するスコアがどちらも 0 となる点が挙げられる．そこで，入力文が改善されれば正の値，悪化すれば負の値をとる尺度である I-measure が提案された．I-measure は入力文，訂正文，リファレンスに対してトークンレベルでアライメントを行い，精度（accuracy）に基づきスコアを計算する．

2.1.3 GLEU+

GLEU+ は機械翻訳の標準的な評価尺度である BLEU [19] を GEC のために改善した評価尺度である．GLEU+ は訂正文 (H) とリファレンス (R) で一致する n-gram 数から，原文 (S) に現れるがリファレンスに現れない n-gram 数を減算することによって計算される．形式的には次式で表される．

$$\mathrm{GLEU+} = \mathrm{BP} \cdot \exp(\sum_{n=1}^{4} \frac{1}{n} \log(p_n\prime)) \tag{1}$$

$$p_n\prime = \frac{N(H,R) - [N(H,S) - N(H,S,R)]}{N(H)} \tag{2}$$

ただし，$N(A,B,C,...)$ は集合間での n-gram 重なり数を表し，BP は BLEU と同様の brave penalty を表す．brave penalty は入力文に対して出力文が短い場合に n-gram 適合率を減点する項である．

2.2 リファレンスレス手法

リファレンスベースの大きな欠点の一つに，リファレンスにない訂正をうまく評価できない問題がある．GEC タスクは機械翻訳や自動要約と同様生成タスクであるため，正解の訂正は一つとは限らない．例えば，次のような例を考える．

(1) a. From this scope, social media has shorten our distance.

b. From this scope, social media has *shortened our distance*.

c. From this *perspective*, social media has *made the world smaller*.

文 (1a) には，(1b) と (1c) のようは訂正例が考えられる．リファレンスとして (1b) しか無い場合，リファレンスベースでは (1c) のような訂正を適切に評価できない．考えられる訂正をリファレンスとして作れば良いが，作成のコストもかかるため妥当な訂正を網羅することは難しい．この

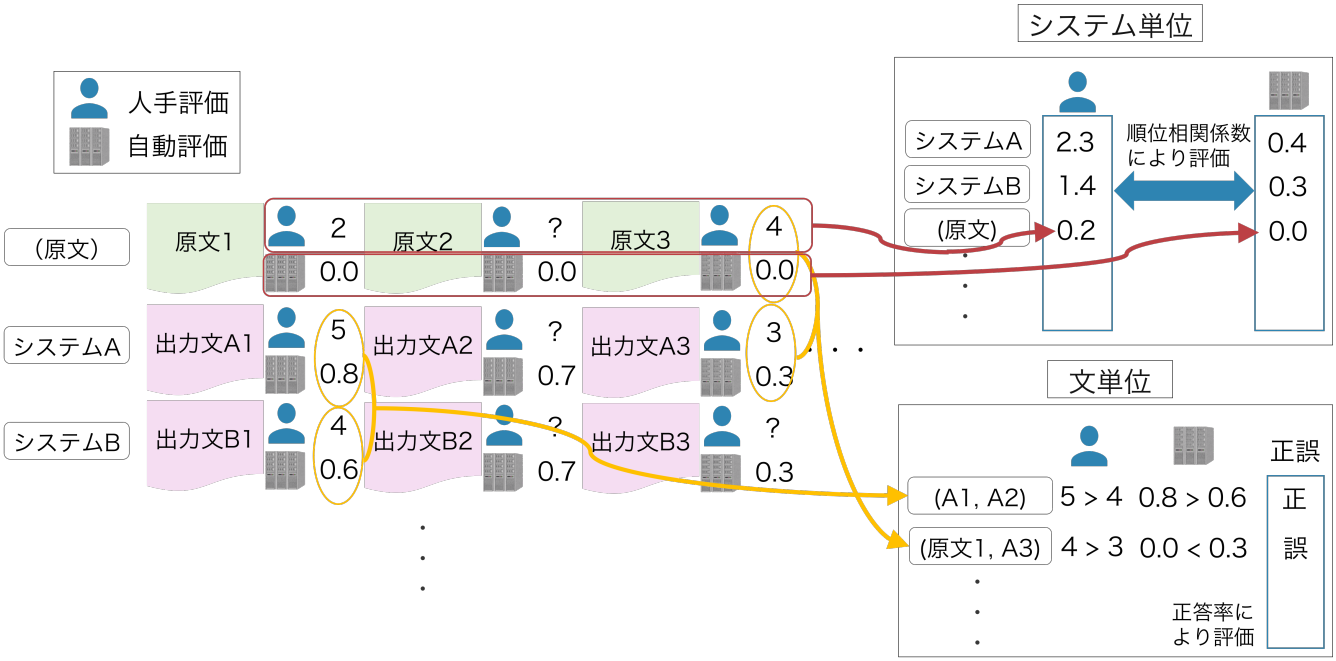

図 1 自動評価尺度のの検証方法.

ような問題を解決するため，リファレンスレス評価が提案された [16].

本節では本稿で用いる Asano らによって提案された手法 [13] について説明する．ある入力文 s に対する訂正文が h であったとき (s,h) に対するスコアを，文法性のスコア S_G，自然さのスコア S_F，文意の保存のスコア S_M の重み付き和によって求める．

$$\mathrm{Score}(h,s) = \alpha \mathrm{S_G}(h) + \beta \mathrm{S_F}(h) + \gamma \mathrm{S_M}(h,s), \qquad (3)$$

ただし $\mathrm{S_G}$, $\mathrm{S_F}$, $\mathrm{S_M}$ の値域は [0,1] であり，$\alpha+\beta+\gamma=1$ である．システムのスコアは各 Score(h, s) の平均を用いる．各観点はリファレンスを用いずに以下の手法によりモデル化する．

2.2.1 文法性

文法性については，与えられた文が文法的である確率をロジスティック回帰により算出して $\mathrm{S_G}(h)$ とする．素性については，Heilman ら [11] が用いたスペルミス数，言語モデルスコア，OOV 数，PCFG およびリンク文法に基づく素性に加え，文法誤り検出数や数の不一致素性などを用いた．Heilaman モデルは Napoles らによる実装[*1]を用いて，Heilman らの GUG データセットで訓練した．素性に用いる言語モデルの学習には Gigaword と TOEFL11 を用いた．

2.2.2 自然さ

自然さは文の出現頻度に左右されることが知られている．本稿では Lau ら [14] と同様に，自然さを次式で算出する[*2].

$$\mathrm{S_F}(h) = \frac{\log P_m(h) - \log P_n(h)}{|h|} \qquad (4)$$

$|h|$ は文長，P_m は言語モデルによる生成確率，P_n はユニグラム生成確率である．言語モデルによる文の生成確率は文長が長いときや希少語が出現するときに低下するが，それは必ずしも自然さの低下を意味しない．そのため文の生成確率を文長とユニグラム生成確率で正規化している．

言語モデルは RNN 言語モデル（実装は faster-rnnlm[*3]）を採用し，全単語を小文字化した British National Corpus[1] と Wikipedia の合計 1000 万文で訓練した．

2.2.3 文意の保存

単純に文意の保存を評価するためには原文と訂正後の文の単語がどのくらい一致しているかを考慮すれば良い．しかし学習者の文で機能語は訂正されることが多く，内容語も活用形や類義語に訂正される場合がある．そこで，学習者の文中の内容語が全く無関係な別の語に訂正されると文意が変わることが多いと仮定する．本稿では訂正前後の文に METEOR 1.5 [8] を適用した．METEOR は本来，機械翻訳の評価ツールであり，システムの出力とリファレンスに対して活用形や類義語を考慮した単語アライメントを行うことでスコアを算出するものである．GEC において訂正前後の文意の保存を評価するために，次式によってスコアを求める．

[*1] https://github.com/cnap/grammaticality-metrics/tree/master/heilman-et-al

[*2] S_N は多くの場合 0 以上 1 未満であるが，0 未満のとき $S_N=0$, 1 以上のとき $S_N=1$ とする

[*3] https://github.com/yandex/faster-rnnlm

評価尺度	Spearman's ρ
M^2	0.648
I-measure	0.769
GLEU+	0.857
Asano et al.(2017)	0.874

表 1　システム単位自動評価と人手評価の順位相関係数

$$P = \frac{m(h_c)}{|h_c|} \tag{5}$$

$$R = \frac{m(s_c)}{|s_c|} \tag{6}$$

$$\mathrm{S_M(h,s)} = \frac{P \cdot R}{t \cdot P + (1-t) \cdot R} \tag{7}$$

h_c は GEC システムの出力中の内容語，s_c は原文中の内容語である．$m(h_c)$ は出力中の内容語のうちアライメントされた単語数，$m(s_c)$ は原文中の内容語でアライメントされた単語数を表す．t の値はデフォルト値である 0.85 を用いた．

2.3　自動評価尺度の評価性能の検証方法

システム単位における自動評価尺度の性能は，人手による評価と比較することで検証されてきた．これをせつめいするために図 1 に例を示す．各 GEC システムに対して人手評価スコアが与えられている．人手評価スコアは，人手で文ごとに評価した少量のデータを使い，レーティングアルゴリズムである TrueSkill [12] を用いて算出される．また，各 GEC システムの出力文は自動評価尺度によって評価されており，各 GEC システムの自動評価スコアは出力文の平均によって計算される．自動評価尺度の良さは，この人手スコアと自動評価スコアの相関係数を比較することで検証されてきた [10], [13], [16], [20].

表 1 は，2.1 節および 2.2 節で説明した自動評価尺度のシステム単位評価の結果である．システム単位評価では，リファレンスレス評価尺度がリファレンスを使うリファレンスベース評価尺度よりも人手に近い評価ができている．

3.　文単位評価の性能調査

これまでの研究で，GLEU+ およびリファレンスレス手法はシステム単位では人手評価と強く相関していることが示されている．しかし，システム単位評価が適切であるからといって，それぞれに文に対して正しくスコアがつけられているとは限らない．例えば，図 2 のような例を考える．この例の人手評価では，システム A が B よりも良いと判断している．システム単位の評価を見ると，自動評価尺度も A に対して 0.8，B に対して 0.6 をつけている．人手評価と同じ結果であり，システム単位では正しく評価ができている．しかし，各文のスコアを見たとき，ある文については人間は A の方が優れていると評価しているにも関わらず，自動評価尺度が B の方が優れていると評価していれば，自動評価尺度は文単位では訂正文を正しく評価できていない．そこで本稿では，これまで提案された自動評価尺度である M^2，I-measure，GLEU+ およびリファレンスレス評価尺度が文単位でどの程度正確に評価できるかを検証する．

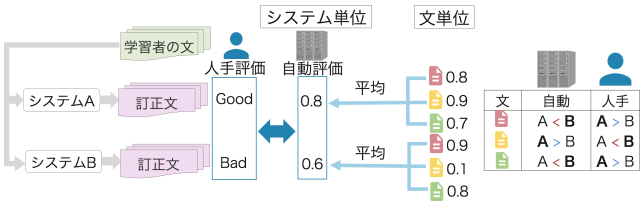

図 2　文単位評価が不適切な例

入力文	
s: Genetic diseases *costs* highly for *the* treatment.	
訂正文	人手評価
h_1: Genetic diseases cost highly for treatment.	1
h_2: Genetic diseases cost *higher* for *the* treatment.	5
h_3: Genetic diseases cost *high* for *the* treatment.	3
h_4: Genetic diseases *costs* highly for treatment.	3

表 2　入力文 s に対する複数の訂正システムの出力 h と人手評価．5 が最も良く，1 が最も悪い．

3.1　文単位評価の性能調査のためのデータ

文単位評価の性能調査のためには，訂正システムの出力それぞれに対して人手評価が付与されているデータが必要である．本研究では，Grundkiewicz ら [10] によって作られたデータおよび Napoles ら [17] によって作られた二つのデータを使用する．これらのデータは本稿で行うような文単位での性能を調査するためではなく，2.3 節で説明したシステム単位の人手評価スコアを計算するために作られた．彼らが作成したデータでは表 2 のように，一つの入力文に対して複数システムの出力が与えられており，それらに対して人手評価が 5 段階の相対評価で与えられている．

Grundkiewicz らのデータは，文法誤り訂正のコンペティションである CoNLL 2014 Shared Task on GEC [18] のテストデータおよび Shared Task 参加システムの出力の一部に人手評価を付与したものである．Napoles らのデータは GUG データセット [11] に対して四つの訂正システムを適用し，その出力に対して人手評価を付与したものである．

3.2　実験設定

リファレンスベースの評価手法に用いるリファレンスは 1 つだけでなく複数用いることができる．本稿では，先行研究 [13], [16] 同様以下のリファレンスを使用した．リファレンスベース手法のリファレンスには，CoNLL 2014 Shared Task on GEC のテストセットのリファレンスを 2 セット，Bryant ら [4] が作成したリファレンスを 8 セット，Sakaguchi らが作成したリファレンスを 8 セット，計

18 セットを用いた．

2.2 節で説明したようにリファレンスレス評価手法では文法性，自然さ，文意の保存の重みを決定する必要がある．本稿では，リファレンスレス手法の計算式 3 における α, β, γ の値は，Asano et al. (2017) と同様に $\gamma = 0.1$ に固定し，システム単位の人手評価との相関係数が最大になるように α, β の値を調整した．Grundkiewicz らのデータでテストする際は Napoles らのデータでチューニングを行い，Napoles らのデータでテストする際は Grundkiewicz らのデータでチューニングを行った．

3.3 文単位評価の良さの検証方法

文法誤り訂正の評価尺度のシステム単位での性能を検証する場合には相関係数が用いられる．しかしながら相関係数は複数システムの出力に対する人手評価が全て同じ，もしくは自動評価が全て同じ値の場合に定義することができない．文単位の場合，自動評価尺度によっては全て同じスコアになる場合があるため，相関係数では適切に評価できない．そこで，本研究では任意の 2 つの訂正に対する人手評価が異なる場合と同じ場合に分けて評価した．

人手評価が異なるペアに対しては，自動評価尺度が人手評価で優れている方に高いスコアが与えられていれば正答とみなし，正答率により評価した．

$$Accuracy = \frac{\text{大小関係を適切に評価できたペア数}}{\text{人手評価の順位が異なるペア数}} \quad (8)$$

例えば，表 2 の例では，(h_1, h_2), (h_1, h_3), (h_1, h_4), (h_2, h_3), (h_2, h_4) の五つの組み合わせが人手評価が異なるペアである．この中の二つが大小関係を適切に判定できている場合は，$Accuracy = 2/5$ になる．人手評価が異なるペアは Grundkiewicz で 14,822 組，Napoles で 608 組存在した．この評価を優劣判定調査と呼ぶ．

人手評価が同じペアは自動評価スコアもできるだけ近い値になるのが望ましい．そのため自動評価スコア同士の平均絶対誤差 (Mean Absolute Error; MAE) で評価した．

$$MAE = \frac{\sum |score_1 - score_2|}{\text{人手評価が同順のペア数}} \quad (9)$$

ただし，もともとスコアの分散が小さい評価尺度が有利になるのを防ぐため，各評価尺度のスコアは平均が 0，分散が 1 になるよう標準化を行った．例えば表 2 における (h_3, h_4) が人手評価が同じペアであり，この二つに対して自動評価尺度で付けたスコアから MAE を計算する．人手評価が同じペアは Grundkiewicz で 5,964 組，Napoles で 64 組存在した．この評価を類似性判定調査と呼ぶ．

3.4 結果

優劣判定調査の結果　人手評価が異なる 2 文に対する優劣判定の正答率を表 3 に示す．リファレンスレス手法 (Asano et al. (2017)) はリファレンスベース手法と比べて高い正答率を示した．リファレンスベース手法の中では GLEU+ が M^2 や I-measure よりも正答率が高かった．Napoles らのデータにおいて GLEU+ はリファレンスレスと同程度（正答数の差は 7）の正答率を示した．

評価尺度	Grundkiewicz	Napoles
M^2	0.594	0.632
I-measure	0.673	0.618
GLEU+	0.675	0.766
Asano et al.(2017)	**0.706**	**0.778**

表 3　人手評価が異なる 2 文に対する優劣判定の正答率

評価尺度	Grundkiewicz	Napoles
M^2	0.919	0.668
I-measure	0.718	0.618
GLEU+	0.429	0.437
Asano et al.(2017)	**0.387**	**0.264**

表 4　人手評価が同じ 2 文に対するスコアの平均絶対誤差

類似性判定調査の結果　人手評価が同じ 2 文に対するスコアの平均絶対誤差を表 4 に示す．リファレンスレス手法の平均絶対誤差が小さく，人手評価が同じ 2 文に対して最も近いスコアを与えることができている．リファレンスベース評価手法の中では，GLEU+ が最も良い結果となっており，優劣判定調査・類似性判定調査の両方で優れている．

表 1 に示したシステム単位評価の結果と文単位評価の結果を比較すると，各評価尺度の性能の序列は文単位でも同じとなっている．しかし，システム単位評価では I-measure と GLEU+ の間に差があるが，優劣判定能力においては差は認められない．一方，類似性判定調査の結果では GLEU+ が I-measure を上回っている．これらの結果からも I-measure は優劣判定はできるが，その評価スコア自体は適切につけられていないことがわかる．

3.5 事例分析

リファレンスレス手法が人手評価が異なる訂正を適切に評価できていた例を示す．表 5 の例で訂正 A は文法的であるが訂正 B は主語と述語の数が一致していないため文法的ではない．この例でリファレンスレス手法は A の方を高く評価できたが，リファレンスベース手法は B の方を高く評価した．これは訂正 B の表層がリファレンスと似ているからであるが，リファレンスベース手法は訂正とリファレンスが異なっている箇所の重大性を考慮せずに評価するからであると考えられる．

一方，リファレンスレス手法は失敗したが従来手法は正答できたものとしては，冠詞だけが異なっている事例が多く見られた．例えば，表 6 における訂正 A には冠詞誤りが二箇所存在する．これは適切な冠詞選択のためには文脈情報が必要なことが多く，リファレンスレス手法は文脈情報を一切用いないのに対し，従来手法は文脈を考慮して作成されたリファレンスと訂正を比較しているからであると考

原文					
On the other hand, the viewers, are not the listeners.					
リファレンス					
On the other hand, the viewers are not the listeners.					
訂正文 A	On the other hand, the viewer, is not the listener.				
	人手	Asano	M^2	IM	GLEU
	3	0.822	0.00	-0.391	0.414
訂正文 B	On the other hand, viewer *are* not listeners.				
	人手	Asano	M^2	IM	GLEU
	2	0.645	0.714	-0.096	0.496

表 5 リファレンスベース手法の優劣判定の誤り例．人手評価は 5 が最も良く，1 が最も悪い．

原文					
In the view of my point , a carrier of a known genetic risk should not be obligated to tell his or her relatives.					
リファレンス					
In my point of view, a carrier of a known genetic risk should not be obligated to tell his or her relatives.					
訂正文 A	In view of my point, *the* carrier of ϕ known genetic risk should not be obligated to tell his or her relatives.				
	人手	Asano	M^2	IM	GLEU
	4	0.763	0.476	-0.789	0.269
訂正文 B	In view of my point, a carrier of a known genetic risk should not be obligated to tell his or her relatives.				
	人手	Asano	M^2	IM	GLEU
	5	0.753	0.625	0.222	0.348

表 6 リファレンスレス手法の優劣判定の誤り例．人手評価は 5 が最も良く，1 が最も悪い．

原文				
With the improvements of technology, a new life with genetic risk can be detected.				
リファレンス				
With the improvements *in* technology, a new life with genetic risk can be detected.				
訂正文 A	With the *improvement* of technology, a new life with genetic risk can be detected.			
	Asano	M^2	IM	GLEU
	0.809	0.0	-0.114	0.449
訂正文 B	With the improvements *in* technology, a new life with genetic risk can be detected.			
	Asano	M^2	IM	GLEU
	0.791	1.0	1.0	0.566

表 7 人手評価が同じ文に対するリファレンスベース手法の誤り例．自動評価スコアは標準化前の値．この 2 文に対する人手評価はともに 4 である．

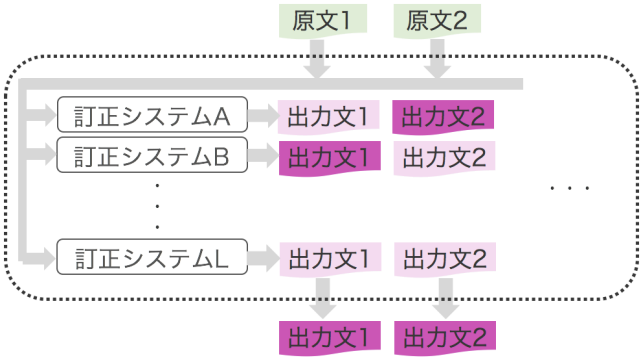

図 3 アンサンブルシステム

える．

人手評価が同じ訂正に対し，リファレンスベース手法の絶対誤差が大きかった例を表 7 に示す．訂正 A と B は人手評価に影響を与えるほどの差異は無い．しかし訂正 A はリファレンスに無く，訂正 B はリファレンスと完全に一致している．このため M^2 および I-measure は人手評価が同じにも関わらず大きく異なる評価を行っている．GLEU+ は比較的近い値をつけている．理由としては，GLEU+ は n-gram 適合率に基づく評価である点や，リファレンスが複数あるときにその平均値を採用している点が考えられる．しかし，標準化を行うとその差は 0.674 となる．一方，リファレンスレス手法は標準化を行ってもその差は 0.109 に収まっており，人間に近い評価ができている．

4. リファレンスレス評価の文法誤り訂正への応用可能性の調査

3 節の実験より，リファレンスレス評価がリファレンスベース評価よりも文単位の評価能力でも優れていることが明らかになった．それを受け，本節ではリファレンスレス評価尺度のもう一つの可能性を調査する．リファレンスレス評価尺度は正解データを必要としないため，正解データのない文に対しても評価スコアを与えることができる．つまり，リファレンスレス評価尺度を使えば，GEC システムの出力した訂正文の候補の中から最もよい訂正文を選択することで誤り訂正ができる可能性がある．そこで最もよい訂正を選択する訂正システムを想定したときに実際に訂正性能が向上するかどうかを調べた．以下，この手法をアンサンブルシステムと呼ぶ．

4.1 リファレンスレス評価を使った文法誤り訂正

図 3 のように，」」各入力文に対する複数の GEC 訂正システムの出力をリファレンスレス手法で評価し，最もスコアの高い訂正を選択するシステムを構築した．評価用のデータとして CoNLL 2014 Shared Task on GEC のテストセットを使用した．アンサンブルするシステムは，CoNLL 2014 Shared Task on GEC 参加 12 システムの訂正結果が公開されているためそれを使用する．[*4]．

[*4] http://www.comp.nus.edu.sg/~nlp/conll14st/official_submissions.tar.gz

評価尺度	アンサンブル	トップシステム
TrueSkill	0.451	0.213
M^2	0.406	0.373
GLEU+	0.551	0.531

表 8 訂正システムに対するスコア．トップシステムは CoNLL2014 参加システムで各スコアが最良のシステムを意味する．

4.2 評価方法

訂正システムの性能が向上するかどうかを調べるために，Grundkiewicz [10] らや Napoles ら [17] がシステム単位の人手評価をするために使った方法を使用する．彼らと同様に，システム単位の人手評価を Grundkiewicz らのデータセットを用いて各システムに対する人手評価を TrueSkill により再計算することにより求めた．ただし，人手評価は一部の入力文（1312 文中 663 文）に対する一部の訂正にしか与えられていないため，アンサンブルシステムは人手評価が与えられている訂正のみを使用した．

また，全入力文に対する訂正を評価するために，リファレンスベース手法による評価も行った．評価尺度としては M^2 と GLEU+ を用いた．リファレンスに 18 セット全てを用い，各文に対するスコアの平均値をシステムのスコアとした．

4.3 結果

アンサンブルシステムによる文法誤り訂正の実験結果を表 8 に示す．いずれの評価でもリファレンスレス手法で訂正を選択することにより訂正性能が向上する結果となった．TrueSkill のスコアが約 2 倍になっていることは訂正が 2 倍改善したことを意味するものは無いが，明らかな性能向上を示している．M^2 スコアや GLEU+についても性能が改善することが確かめられた．

この実験結果からリファレンスレス評価手法は，文法誤り訂正の性能向上に有用であると言える．また，本稿で行ったアンサンブル手法ではなく，リファレンスレス評価手法のコンポーネントである文法性，自然さ，文意の保存の尺度を直接 GEC システムの中に取り込んだモデルを作ることも考えることができる．

5. 関連研究

機械翻訳の自動評価の分野では BLEU や METEOR といったリファレンスベース手法が提案されている．メタ評価には，システム単位ではピアソンの相関係数，文単位では同順を無視したケンドールの順位相関係数が用いられている [3]．本稿では，平均絶対誤差によって同順の訂正に対しても同じような評価できているかを調査した．

機械翻訳の分野では，リファレンスを用いずに翻訳を評価する品質推定（Quality Estimation）と呼ばれるタスクも行われている．この分野では一貫した人手評価が各文に与えられているデータセットが作成されているため，評価にはピアソンの相関係数およびスピアマンの順位相関係数が用いられる [2]．一方，文法誤り訂正の分野で評価尺度の良さを文単位で検証する研究は本稿が初である．

6. おわりに

文法誤り訂正の自動評価尺度の性能評価はこれまでシステム単位評価で行われてきており，訂正文ごとにスコアが適切に付けられているかは調査されていなかった．そこで本稿では，文法誤り訂正の自動評価尺度の性能評価に置いて，初めて文単位での性能評価を行った．文単位評価を優劣判定と類似性判定という二つの観点に分けて調べた結果，文単位評価においてもリファレンスレス評価尺度がリファレンスベース評価尺度より優れていることを明らかにした．また，リファレンスレス評価を使ったアンサンブル手法による誤り訂正の性能を調査し，リファレンスレス評価尺度を使うことで文法誤り訂正の性能を向上させることができることを明らかにした．

参考文献

[1] BNC Consortium: *The British National Corpus*, version 3 (BNC XML Edition), Distributed by Oxford University Computing Services on behalf of the BNC Consortium (2007).

[2] Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K. and Zampieri, M.: Findings of the 2016 Conference on Machine Translation, *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, Association for Computational Linguistics, pp. 131–198 (2016).

[3] Bojar, O., Graham, Y. and Kamran, A.: Results of the WMT17 Metrics Shared Task, *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 489–513 (2017).

[4] Bryant, C. and Ng, H. T.: How Far are We from Fully Automatic High Quality Grammatical Error Correction?, *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 Paper*, pp. 697–707 (2015).

[5] Dahlmeier, D. and Ng, H. T.: Better evaluation for grammatical error correction, *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 568–572 (2012).

[6] Dale, R., Anisimoff, I. and Narroway, G.: HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task, *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, Montréal, Canada, Association for Computational Linguistics, pp. 54–62 (2012).

[7] Dale, R. and Kilgarriff, A.: Helping Our Own: The HOO 2011 Pilot Shared Task, *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, Nancy, France, Associa-

tion for Computational Linguistics, pp. 242–249 (2011).
[8] Denkowski, M. and Lavie, A.: Meteor Universal: Language Specific Translation Evaluation for Any Target Language, *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380 (2014).
[9] Felice, M. and Briscoe, T.: Towards a standard evaluation method for grammatical error detection and correction, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 578–587 (2015).
[10] Grundkiewicz, R., Junczys-Dowmunt, M. and Gillian, E.: Human Evaluation of Grammatical Error Correction Systems, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 461–470 (2015).
[11] Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M. and Tetreault, J.: Predicting Grammaticality on an Ordinal Scale, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 174–180 (2014).
[12] Herbrich, R., Minka, T. and Graepel, T.: TrueSkill™: A Bayesian Skill Rating System, *Advances in Neural Information Processing Systems 19* (Schölkopf, B., Platt, J. C. and Hoffman, T., eds.), MIT Press, pp. 569–576 (2007).
[13] Hiroki, A., Tomoya, M. and Kentaro, I.: Reference-based Metrics can be Replaced with Reference-less Metrics in Evaluating Grammatical Error Correction Systems, *The 8th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics (2017).
[14] Lau, J. H., Clark, A. and Lappin, S.: Unsupervised Prediction of Acceptability Judgements, *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. 1618–1628 (2015).
[15] Napoles, C., Sakaguchi, K., Post, M. and Tetreault, J.: Ground Truth for Grammatical Error Correction Metrics, *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 (2015).
[16] Napoles, C., Sakaguchi, K. and Tetreault, J.: There's No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2109–2115 (2016).
[17] Napoles, C., Sakaguchi, K. and Tetreault, J.: JFLEG: A Fluency Corpus and B enchmark for Grammatical Error Correction, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 229–234 (2017).
[18] Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H. and Bryant, C.: The CoNLL-2014 Shared Task on Grammatical Error Correction, *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–14 (2014).
[19] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation, *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318 (2002).
[20] Sakaguchi, K., Napoles, C., Post, M. and Tetreault, J.: Reassessing the Goals of Grammatical Error Correction: Fluency Instead of Grammaticality, *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 169–182 (2016).