B7IM2005

修士論文

流暢性・意味保存性を考慮したニューラル文法誤り訂正

浅野 広樹

2019 年 2 月 5 日

東北大学 大学院
情報科学研究科 システム情報科学専攻

本論文は東北大学 大学院情報科学研究科 システム情報科学専攻に
修士 (情報科学) 授与の要件として提出した修士論文である。

浅野 広樹

審査委員：
乾 健太郎 教授 （主指導教員）
張山 昌論 教授
篠原 歩 教授
鈴木 潤 准教授 （副指導教員）

流暢性・意味保存性を考慮したニューラル文法誤り訂正*

浅野 広樹

内容梗概

文法誤り訂正（Grammatical Error Correction:GEC）は，言語学習者の書いた文の文法的な誤りを訂正するタスクである．このタスクのゴールは自動で人間らしい訂正ができるようになることである．訂正システムの研究開発段階においては，システムの性能が改善したかどうかを高速で評価する必要があり，このときの自動評価手法も人間らしい評価であることが望ましい．そこで，本論文でははじめに（1）文法誤り訂正の自動評価手法に関する研究について報告する．この研究では，訂正の質を正解データを用いずに自動評価する手法を提案し，従来の正解データを用いる手法よりも正確な評価ができる可能性を示す．次に，(2) 訂正システムの性能向上を目指す研究について報告する．文法誤り訂正においては一度の訂正で全ての誤りを訂正するのが困難な事例があることから，訂正を反復することで性能向上を図る．訂正の反復による効果を調べる実験を行い，訂正システムの今後の課題について考察する．

キーワード

自然言語処理，深層学習，文法誤り訂正，自動評価尺度

*東北大学 大学院情報科学研究科 システム情報科学専攻 修士論文, B7IM2005, 2019 年 2 月 5 日.

Contents

List of Figures

List of Tables

1　はじめに

文法誤り訂正（Grammatical Error Correction:GEC）は，言語学習者の書いた文の文法的な誤りを訂正するタスクである．このタスクのゴールは自動で人間らしい訂正ができるようになることである．訂正システムの研究開発段階においては，システムの性能が改善したかどうかを高速で評価する必要があり，このときの自動評価手法も人間らしい評価になっていることが望ましい．そこで，本論文でははじめに文法誤り訂正の自動評価手法に関する研究について述べる．次に，訂正システムの性能向上を目指した，新たな訂正システムに関する研究について述べる．

2　文法性・流暢性・意味保存性に基づく文法誤り訂正の参照無し評価

2.1　背景

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るため，与えられた入力に対する出力の正解が 1 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 システムの性能評価には，システムの出力を正解データ（参照文）と比較することにより評価する手法（参照有り手法）が一般的に用いられている．この参照有り手法では，訂正が正しくても参照文に無ければ減点されるため，正確な評価のためには可能な訂正を網羅する必要がある．しかし参照文の作成は人手で行う必要があるためコストが高く，可能な訂正を全て網羅することは現実的ではない．この問題に対処するため，[1] は参照文を使わず訂正文の文法性に基づき訂正を評価する手法を提案した．しかし参照有り手法である GLEU [2] を上回る性能での評価は実現できなかった．

そこで本論文では [1] の参照無し手法を拡張し，その評価性能を調べる．具体的には，[1] が用いた文法性の観点に加え，流暢性と意味保存性の 3 観点を考慮する組み合わせ手法を提案する．流暢性は GEC システムの出力が英文としてどの程度自然であるかという観点であり，意味保存性は訂正前後で文意がどの程度保たれているかという観点である．各評価手法により訂正システムの性能の評価を行ったところ，提案手法が参照有り手法である GLEU よりも人手評価と高い相関

を示した．

これに加えて，各自動評価尺度の文単位での評価性能を調べる実験も行った．文単位での評価が適切にできれば，GEC システムの人手による誤り分析に有用であるが文法誤り訂正の自動評価において文単位の性能を調べた研究はこれまでない．そこで，文単位評価の性能を調べる実験を行ったところ，提案した参照無し手法が参照有り手法より高い性能を示した．この結果を受けて，参照無し手法のもうの可能性も調査した．参照無し手法は正解を使わずに与えられた文を評価できるため，複数の訂正候補の中から最も良い訂正文を選択するために本手法が使えると考えられる．このことを実験的に確かめるために複数の GEC システムの出力を参照無しで評価し，最も良いものを採用するアンサンブル手法の誤り訂正性能を調べたところ，アンサンブル前のシステムの性能を上回った．

2.2 自動評価尺度の評価方法

自動評価尺度に求められる性質のうち最も重要なものは，人手評価との相関が高いことであるとされている [3]．このため，評価尺度の良さは人手との相関係数で評価されるのが一般的である．機械翻訳の評価尺度の shared task である WMT 2017 Metrics Shared Task [4] においても自動評価尺度は人手評価との相関によって比較されている．このタスクにおいて評価尺度のメタ評価には，翻訳システム単位と文単位で評価が行われている．システム単位のメタ評価では，人手評価によるシステムに対する評価と自動評価尺度によるシステムに対する評価を比べることで評価する．文単位のメタ評価では，システムの翻訳ごとに人手で優劣が付けられており，自動評価尺度によってその優劣を識別できるかで評価する．システム単位の評価尺度に対してはピアソンの相関係数やスピアマンの順位相関係数，文単位の評価尺度に対してはケンドールの順位相関係数[1]が用いられた．

文法誤り訂正の分野においても自動評価尺度の性能は，訂正システムに対する人手評価スコアと自動評価スコアの相関によって検証されてきた [5, 2, 1]．一方で，我々の知る限り，自動評価尺度の文単位での性能は検証されていない．そこで本研究では，提案手法と従来手法の自動評価尺度を先行研究に従ってシステム単位で比較するとともに，機械翻訳タスクで行われているように各評価尺度の文単位評価における性能も調査する．システム単位評価，文単位評価に関しては 2.5.1 節，2.5.2 節でそれぞれ詳述する．

[1]WMT 2017 Metrics Shared Task では人手評価で同順とされた文対を除外する計算法が使用された．

2.3　既存の評価尺度

機械翻訳の分野では，BLEU [6] などの自動評価尺度によって翻訳システムが比較できるようになり，研究が発展してきた．文法誤り訂正の分野においても自動評価尺度は重要である．これまでの文法誤り訂正の研究では，機械翻訳と同様に参照有り手法による自動評価が用いられてきた．そこで本節では参照有り評価尺度の代表的な手法について述べ，その後参照無し評価尺度の手法について述べる．

2.3.1　参照有り手法

訂正システムの評価では，学習者の書いた文に対する訂正の正解データ（参照文）を使うことが一般的である．この参照有り評価は M^2 [7]，I-measure [8]，GLEU [9, 2] が考案されている．参照有り手法では正確な評価のために，各入力文に対する参照文を 1 個だけでなく複数個用いることができる．参照文を複数用いる場合，各文の評価は M^2 および I-measure では最大値が採用され，GLEU は平均値が採用される．

M^2 文法誤り訂正の初期の研究では，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 [10, 11]．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．この問題を解決するために 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．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．

I-measure 上述の M^2 の問題点として，訂正を全く行わないシステムと誤った訂正をした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，参照文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．

GLEU 機械翻訳の標準的な評価尺度である BLEU [6] を GEC のために改善した評価尺度である．GLEU は訂正文 (H) と参照文 (R) で一致する n-gram 数から，原文 (S) に現れるが参照文に現れない 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．

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

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

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．これまでに提案された参照有り手法の中では最も人手評価との相関が高い [1]．

2.3.2 参照無し手法

機械翻訳の分野では，参照文を用いずに翻訳の品質を評価する品質推定（Quality Estimation）と呼ばれるタスクも行われており，近年は shared task も開催されている [12, 13]．機械翻訳の品質推定タスクでは，翻訳システムの出力の良さを測るために，Human-targetd Translation Error Rate (HTER) [14] と呼ばれる，人間の翻訳とシステムの翻訳の編集距離がどの程度近いかを計算する指標が用いられる．機械翻訳の品質推定の手法では，各システムの出力に対して HTER が付与された大量のデータを用いてシステムを学習する．文法誤り訂正の参照無し評価用のデータセットには，一部の少量の文に対してのみ人手の評価が付与されているため，品質推定の手法を文法誤り訂正の参照無し評価に応用することは難しい．

文法誤り訂正の分野では，参照文を用いずに訂正の品質を評価する手法を [1] が初めて提案した．文法誤り訂正では訂正システムの入出力文に対して訂正の品質が付与されたデータが十分にないため，訳文品質推定の標準的な手法を用いることができない．そこで [1] は，訂正システムの出力文の文法性を評価する３つの手法を提案した．１つ目は e-rater® による文法誤り検出数に基づく評価，２つ目は Language Tool [15] による文法誤り検出数に基づく評価，３つ目は [16] の言語学的な素性に基づく文法性予測モデルを用いた評価である．実験の結果，e-rater® を用いる手法が最も優れており，参照有り手法である GLEU と同等の性能であることが示された．しかし，e-rater® は通常，自然言語処理の研究目的でオープンに利用することはできない[2]．そこで，本研究では e-rater® を用いず，Language Tool および [16] のモデルなどを組み合わせることで性能向上を図る．

[2] e-rater® は English Testing Service (ETS) の作文評価サービス Criterion の１機能として提供されており，Criterion は教育機関向けの有償サービスであるため．

2.4　提案手法

人手評価に近い参照無し評価を実現するために，[1] の文法性に基づく参照無し評価を拡張する．人手による訂正の傾向を捉えるために，文法性，流暢性，意味保存性の 3 つの観点を考慮した参照無し評価手法（各観点，意味保存性 (Meaning preservation)，流暢性 (Fluency)，文法性 (Grammaticality) の頭文字を取って **MFG** と呼ぶ.）を提案する．[17] が参照文作成の際に用いたガイドラインでは，自然な文にすること，文法的な誤りは訂正すること，文意は保存することが指示されており，人手による訂正では一般的にこのような観点に基づいて訂正されることが多い．

文法性は，GEC システムの出力に標準英語上の文法誤りがあるかどうかという観点であり，先行研究でも用いられた [1]．流暢性は，GEC システムの出力がどの程度自然な英文であるかという観点である．この観点は先行研究 [2] において文法性と区別され，重要性が示された．意味保存性は，訂正の前後で文意が変わっていないかという観点である．

提案手法は，参照文を使わずにこれら 3 つの観点に基づき GEC システムを評価する．本稿では，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，流暢性のスコア S_F，意味保存性のスコア S_M の重み付き和によって求める．

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

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 $\mathrm{Score(h, s)}$ の平均を用いる．各観点は参照文を用いずに以下の手法によりモデル化する．

2.4.1　文法性

[1] が参照無し評価に用いたモデルのうち，[16] の言語学的な素性に基いたモデルを行う手法をベースに用いる．具体的には，文法性のスコア $\mathrm{S_G}(h)$ は言語学的な素性に基づくロジスティック回帰により求める．素性については，[16] が用いたスペルミス数，n-gram 言語モデルスコア，out-of-vocabulary 数，PCFG およびリンク文法に基づく素性に加え，依存構造解析に基づく数の不一致素性と Language Tool[3] による誤り検出数を素性として用いた．

[3] `https://languagetool.org`

モデルの学習はGUGデータセット [16] に対して [1] の実装[4]を用いた．さらに，言語モデルの学習のためにGigaword [18] とTOEFL11 [19] を用いた．

GUGデータセットのテストセットにおいて文法性2値予測タスクを行ったところ，元々の [1] 実装の正解率が77.2% だったのに対し，我々が修正を加えたモデルの正解率は78.9%であった．

2.4.2 流暢性

文法誤り訂正における流暢性の重要性は [2, 17] において示されたが，流暢性を考慮する参照無し評価手法はこれまでに提案されていない．流暢性は言語モデルによってとらえることができる [20]．具体的には，訂正文 h に対し，流暢性 $\mathrm{S_F}(h)$ を次のように求める[5]．

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

$|h|$ 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．

本研究では，言語モデルにはRecurrent Neural Network (RNN) 言語モデル [21] を採用し，実装はfaster-rnnlm[6]を用いた．学習にはBritish National Corpus [22] およびWikipediaの1000万文を用いた．作成したモデルは [20] のテストデータにおいて，人間の容認性判断に対するピアソンの相関係数が0.395であった．

2.4.3 意味保存性

文法誤り訂正においては原文の意味が訂正後も保存されていることは重要である．例えば, 以下の文 (1a) が文 (1b) に訂正される事例を考える.

(1) a. *It is unfair to release a law only point to the genetic disorder.* (original)

b. *It is unfair to pass a law.* (revised)

文 (1b) は文法的であるが，文 (1a) の意味が保存されていないため，文 (1b) は不適切な訂正である.

意味がどの程度保存されているかを測る単純な方法は，原文の単語が訂正後の文でも出現する割合を計算する方法である．このような目的のために機械翻訳の

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

[5] 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

[6] `https://github.com/yandex/faster-rnnlm`

評価尺度を用いる方法が考えられる．本研究では，METEOR [23] を訂正前後の文に適用することで，どの程度文意が保存されているかを評価する．METEOR は BLEU などの評価尺度と比べて意味的な類似度を重視した評価尺度である．本稿では入力文 s と訂正文 h に対する意味保存性のスコア $\mathrm{S_M}(h,s)$ を次式により求める．

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

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

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

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マッチングされた単語数，$m(s_c)$ は原文中の内容語でマッチング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．METEOR の単語マッチングでは表層だけでなく，活用形，類義語，パラフレーズも考慮される．これに加え，本稿ではスペルミスが訂正されてもマッチングされるよう，スペルチェッカを用いて METEOR を拡張した．

2.5 実験

2.2 節で述べたように，本研究ではシステム単位と文単位で評価尺度のメタ評価を行うことで参照無し評価の有効性を確かめる．

2.5.1 自動評価尺度による訂正システム単位評価

本節では，提案手法および従来手法による自動評価がシステム単位の評価でどの程度人間に近いかを調べるための実験について述べる．

実験設定 [1] と同様に，各自動評価手法で訂正システムの出力文を評価し，各文に対するスコアの平均を訂正システムに対するスコアとし，図 1 のように人手評価と比較することで評価尺度のよさを調査した．人手評価との近さを測るためにピアソンの相関係数とスピアマンの順位相関係数を用いた．各相関係数は [5] の Table 3c の人手評価を用いて計算した．

この実験では，CoNLL 2014 Shared Task [24] のデータセット，およびそれに対して [5] が作成した人手評価を用いた．このデータセットは，テストデータ 1,312

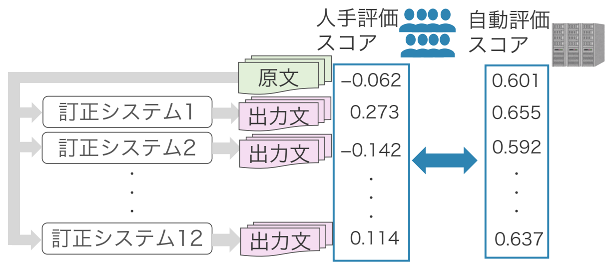
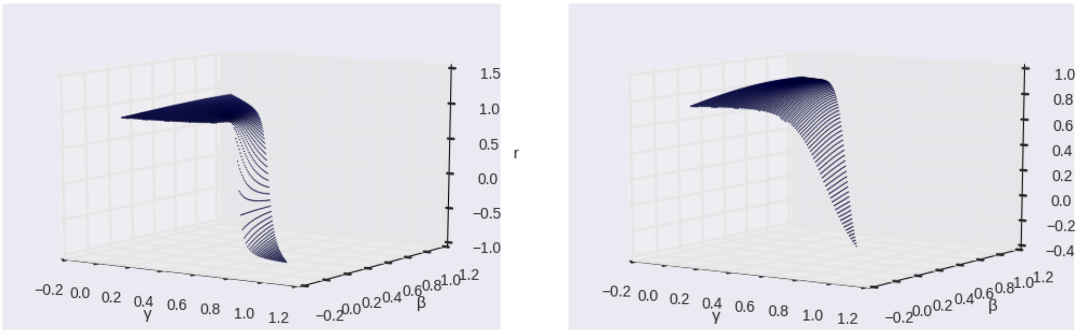


Figure 1: 自動評価尺度のシステム単位評価

文と，それに対する参加 12 システムの訂正結果を含む．このデータに対し，[5] 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 [25] を用いて訂正システム単位の人手評価スコアを算出した．また，このテストデータに対しては多くの参照文が作成されている．公式の参照文が 2 個，[26] による参照文が 8 個，[2] による参照文が 8 個作成されている．本実験では，従来の参照有り手法の性能を最大にするためにこれら 18 個全ての参照文を用いた．

提案手法である MFG の重み α, β, γ の選択は JFLEG データセットを用いて行った．これは CoNLL データセットを dev データと test データに分割することができないためである．また，実際に MFG を使ってシステムを評価する際にも，全く同じシステムの集合に対して人手順位評価がついているデータセットが事前に手に入ることは期待できないため，システム単位の評価では dev データと test データに分割して重みを決めることは適切ではない．GEC の評価尺度の性能評価に使えるデータセットは現在 CoNLL と JFLEG の 2 つしかないため，本研究では JFLEG データセットで重みを調整した．ただし，このデータセットは CoNLL データセットとは次の 2 点において性質が大きく異なる．(1) 訂正システムの数が異なる．CoNLL データセットには 12 システムが含まれているのに対し，JFLEG データセットには 4 システムしか含まれていない．(2) 各システムの編集率の分散が小さい．CoNLL データセットにおいて各システムが訂正した文の割合は 3.7%〜77%なのに対し，JFLEG では 56%〜74%である．このように性質の大きく異なるデータセットを使った場合にも一方で調整した α, β, γ が他方でもうまく働くとすれば，将来においても α, β, γ の調整はそれほど困難にならない可能性がでてくると考えられる．

そこで，性質の大きく異なるデータセットで重みの調整が可能かを調査するために，実際に，CoNLL と JFLEG の 2 つのデータセットにおいて重みの値を 0.01



JFLEG　　　CoNLL

Figure 2: JFLEG データセットと CoNLL データセットにおけるピアソン相関係数．x 軸は γ，y 軸は β, z 軸はピアソンの相関係数を表す．

刻みでグリッドサーチし，ピアソンの相関係数を計算し各データセットの傾向を調査した．図 2 に結果を示す．どちらのデータセットにおいても概ね同じ傾向が見られた．いずれのデータセットにおいても，α, β, γ の値の広い領域で安定的に高い性能を示しており，またその領域は 2 つのデータセットで概ね一致している．このことは，一方のデータセットで調整した α, β, γ がもう一方のデータセットでも有効に働くことを意味している．そこで，JFLEG データセットを使って適当な α, β, γ を選択し，その重みが CoNLL データセットにおいても有効であるかを実験する．具体的には，JFLEG データセットにおいて相関が 0.9 以上となっている α, β, γ の領域（図 2 を z 軸から見た図 3）の中心の点の重み，およびその周辺 4 点の重みを用いた．中心の重みは $(\alpha, \beta, \gamma) = (0.35, 0.35, 0.3)$，周辺の 4 点の重みはそれぞれ $(\alpha, \beta, \gamma) = (0.25, 0.35, 0.4), (0.25, 0.45, 0.3), (0.45, 0.25, 0.3), (0.45, 0.35, 0.2)$ を使用した．

実験結果表 1 に各手法の人手との相関を示す．3 つの観点を用いる提案手法は，中心点の重みを使った場合とその周辺の点 4 つの内，最も高い相関だった点と最も低い相関だった点の結果を示す．文法性のみ，流暢性のみの評価尺度では M^2 を上回ったが GLEU には及ばなかった．意味保存性のみの評価は人手評価との相関が弱いという結果になった．しかし，意味保存性に流暢性を組み合わせることにより性能が改善し，GLEU を上回った．意味保存性，すなわち METEOR は，表層の類似度に基づく評価となっているため，あまり訂正を行わないシステムに対し高い評価を与えてしまう．それにもかかわらず，流暢性と組み合わせたときに重要な役割を果たしていると考えられる．また，3 観点を全て組み合わせるとさらに性能が向上 $(\rho = 0.885)$ した．この結果の意義は，参照無しでも参照有り

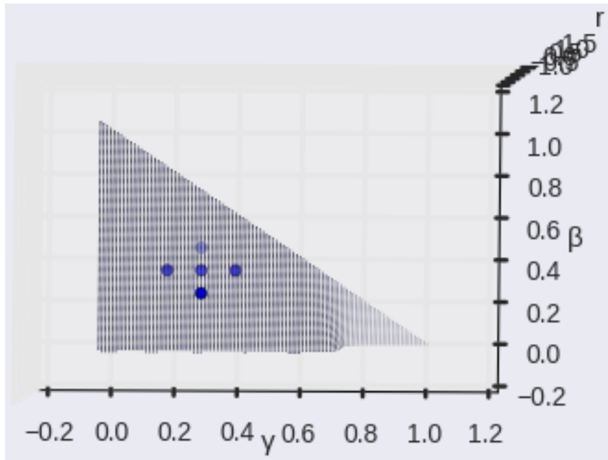


Figure 3: 図 2 左を z 軸方向から見た図

手法である GLEU よりも人手に近い評価ができる可能性を初めて示したことである．また，我々の知る限りこの値は参照無し手法の最高性能である．

また，中心点の周辺の点の重みで実験した結果，相関が最も高い点では $\rho = 0.912$ となり，相関が最も低い点で $\rho = 0.851$ となった．相関が最も低い点でも GLEU とほぼ同等の性能であり，ピアソンの相関係数では GLEU を上回った．この結果は特定の 2 つのデータセットから得られた結果であり，全く新しいデータセットに必ずしも一般化して適応できるわけではないが，性質の異なるデータセットを開発データとして用いたとしても，参照無し手法で参照有り手法を越える可能性があることを実験的に明らかにしたことに意義がある．

一方，本実験では文法性の必要性は示されなかった．本実験では 3 観点から文法性を除いたときの方が高性能（$\rho = 0.929$）となったことからも，文法性は $\alpha, \beta, , \gamma$ の調整次第ではかえって悪影響を与える場合があるといえる．本実験で流暢性モデルとして用いた RNN 言語モデルでは文構造を完全には捉えられないと言われているが [27]，学習者の文の大半は単純な構造であることと，一般に流暢な文は文法的であることが多いことから，流暢性モデルが文法性モデルを包含している部分があり，文法性モデルを用いなくても十分正確な評価ができたと考えられる．流暢性を除いた場合の相関が $\rho = 0.786$，意味保存性をのぞいた場合の相関が $\rho = 0.863$ となり，3 つの観点を使った場合よりも低い相関になっていることから，流暢性・意味保存性は参照無し評価において重要であると言える．

評価尺度	Spearman's ρ	Pearson's r
M^2	0.648	0.632
I-measure	0.769	0.739
GLEU	0.857	0.843
文法性	0.835	0.759
意味保存性	-0.192	0.198
流暢性	0.819	0.864
文法性+意味保存性	0.786	0.771
意味保存性+流暢性	0.929	0.890
流暢性+文法性	0.863	0.844
MFG 中心	0.885	0.878
MFG 最大	0.912	0.898
MFG 最小	0.851	0.854

Table 1: 自動評価による訂正システムのランキングと人手評価間の相関係数.

2.5.2 文単位評価の性能調査

節 2.5.1 の実験で，GLEU および提案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を示した．しかし，システム単位評価が適切であるからといって，それぞれの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 4 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 よりもよいと判断している．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．これは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．文単位で見ると 3 文中 2 つがシステム A がよいと言っているが，自動評価尺度の結果は真逆になっている（図 4 の右）．このように文単位のスコアを見たとき，自動評価による優劣判定が人手評価と異なっている文があれば，その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ことになる．そこで本研究では，これまで提案された自動評価尺度である M^2，I-measure，GLEU および参照無し評価尺度が文単位でどの程度正確に評価できるかを調査する．

文単位評価の実験設定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．本研究では前節で用いたデータ，すなわち [5] によって作られたデータを使用する．このデータはシステム単位の人手評価のために作られたものではあるが，訂正システムの

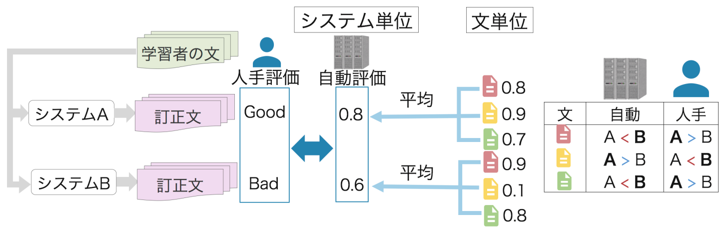


Figure 4: 文単位評価が不適切な例

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

Table 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．

各出力に対して人手評価が付与されているため，その情報を用いる．具体的には表 2 のように，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．

文単位評価の場合，あるテストセットに対する複数システムの出力が得られた時，そのごく一部を人手で評価し，残りを自動評価尺度で評価することは，必ずしも不自然な設定ではない．そこで，本実験では CoNLL データセットを dev データと test データに分割することで提案手法である MFG の重み α, β, γ を調整した．今回は CoNLL データセットをおよそ 1:9 の割合で dev セットと test セットに分割し，dev セット上で後述の正解率が最大となる重みを 0.01 刻みのグリッドサーチにより調整した．調整の結果，$(\alpha, \beta, \gamma) = (0.03, 0.51, 0.46)$ の重みとなった．

文単位評価のメタ評価方法 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を用いた．しかしながら通常の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訂正が異なっていても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．また，人手評価が同じ訂正に関しては自動評価尺度で近いスコアが付くことが望ましいと考えられる．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，正解率（Accuracy）により評価した．

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

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 5 つの組み合わせが人手評価の異なるペアである．自動評価尺度がこのうち 2 つのペアの大小関係を適切に評価できた場合，$Accuracy = 2/5$ になる．また，2.2 節で述べた，WMT17 Metrics Shared Task で使用されたケンドールの順位相関係数 τ による評価も行った．

$$\tau = \frac{\text{大小関係を適切に評価できたペア数} - \text{大小関係を逆順に評価したペア数}}{\text{人手評価の順位が異なるペア数}} \tag{9}$$

この τ は，Accuracy と比べると，人手評価の順位が異なっているにもかかわらず，自動評価で同じ値がつく事例を軽視している．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．

$$MAE = \frac{\sum |score_1 - score_2|}{\text{人手評価が同順のペア数}} \tag{10}$$

例えば表 2 における (h_3, h_4) は人手評価が同じペアであり，この 2 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．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 になるよう標準化した．この評価を類似性判定調査と呼ぶ．

テストデータとして用いる [5] の人手評価は，8 人の評価者がそれぞれ CoNLL2014 Shared Task のデータからサンプリングされた入力文および訂正文に対してランキングを付与することによって作成された．このため，一部の（入力文，訂正文）の組については複数人のランキングが付与されているが，本実験ではそれらを別インスタンスと見なして評価した．テストデータにおいて，優劣判定調査の対象は 14,822 組，類似性判定調査の対象は 5,964 組存在した．

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性能を表 3 に示す．提案手法である MFG は参照有り手法と比べて高い正解率を示した．参

評価尺度	正解率	Kendall's τ
M^2	0.592	0.360
I-measure	0.670	0.390
GLEU	0.671	0.344
MFG	**0.712**	0.425

Table 3: 人手評価が異なる 2 文に対する優劣判定の性能

評価尺度	平均絶対誤差
M^2	0.923
I-measure	0.722
GLEU	0.428
MFG	**0.402**

Table 4: 人手評価が同じ 2 文に対するスコアの平均絶対誤差

照有り手法の中では GLEU が M^2 や I-measure よりも正解率が高かった．MFG と GLEU の正解率の差についてマクネマー検定を行ったところ，5%水準で統計的に有意であった．ケンドールの順位相関係数においても提案手法は参照有り手法よりも高い性能を示した．参照有り手法の中では I-measure が最も高い τ 値を示した．提案手法と I-measure の τ 値の差についてブートストラップ検定を行ったところ，5%水準で統計的に有意であった．

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．MFG 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．参照有り評価手法の中では，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．

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 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示唆される．

事例分析参照無し手法が人手評価の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．この例で参照無し手法は A の方を高く評価できたが，参

<table>
<tr><td colspan="6">原文</td></tr>
<tr><td colspan="6">On the other hand, the viewers, are not the listeners.</td></tr>
<tr><td colspan="6">リファレンス</td></tr>
<tr><td colspan="6">On the other hand, the viewers are not the listeners.</td></tr>
<tr><td rowspan="3">訂正文 A</td><td colspan="5">On the other hand, the viewer, is not the listener.</td></tr>
<tr><td>人手評価</td><td>提案手法</td><td>M^2</td><td>I-measure</td><td>GLEU</td></tr>
<tr><td>3</td><td>0.892</td><td>0.00</td><td>-0.391</td><td>0.414</td></tr>
<tr><td rowspan="3">訂正文 B</td><td colspan="5">On the other hand, viewer are not listeners.</td></tr>
<tr><td>人手評価</td><td>提案手法</td><td>M^2</td><td>I-measure</td><td>GLEU</td></tr>
<tr><td>2</td><td>0.651</td><td>0.714</td><td>-0.096</td><td>0.496</td></tr>
</table>

Table 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．

照有り手法は B の方を高く評価した．これは訂正 B の表層が参照文と似ているからであるが，参照有り手法は訂正と参照文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，参照無し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しており，参照無し評価尺度では人手評価が高い方に低いスコアをつけてしまっている．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，参照無し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参照文と訂正を比較しているからであると考えられる．

人手評価が同じ訂正に対し，参照有り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 は参照文に無く，訂正 B は参照文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，参照文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，参照無し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．

2.5.3　参照無し評価の文法誤り訂正への応用可能性の調査

2.5.2 節の実験より，提案手法が文単位においても参照有り手法を上回る可能性があることが明らかになった．それを受け，本節では参照無し評価尺度のもうの可

<table>
<tr><td colspan="6">原文</td></tr>
<tr><td colspan="6">In the view of my point , a carrier of a known genetic risk should not be obligated to tell his or her relatives.</td></tr>
<tr><td colspan="6">リファレンス</td></tr>
<tr><td colspan="6">In my point of view, a carrier of a known genetic risk should not be obligated to tell his or her relatives.</td></tr>
<tr><td rowspan="3">訂正文 A</td><td colspan="5">In view of my point, *the* carrier of ϕ known genetic risk should not be obligated to tell his or her relatives.</td></tr>
<tr><td>人手評価</td><td>提案手法</td><td>M^2</td><td>I-measure</td><td>GLEU</td></tr>
<tr><td>4</td><td>0.857</td><td>0.476</td><td>-0.789</td><td>0.269</td></tr>
<tr><td rowspan="3">訂正文 B</td><td colspan="5">In view of my point, a carrier of a known genetic risk should not be obligated to tell his or her relatives.</td></tr>
<tr><td>人手評価</td><td>提案手法</td><td>M^2</td><td>I-measure</td><td>GLEU</td></tr>
<tr><td>5</td><td>0.851</td><td>0.625</td><td>0.222</td><td>0.348</td></tr>
</table>

Table 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．

能性を調査する．参照無し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，参照無し評価尺度を使えば，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の精度を改善できる可能性がある．そこで本節では，複数の訂正候補から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．

実験設定図 5 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参照無し手法で評価し，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としては，CoNLL 2014 Shared Task on GEC 参加 12 システムの訂正結果を使用する．[7]

評価方法アンサンブルにより GEC システムの性能が向上するかどうかを調べるために，[5] や [17] がシステム単位の人手評価値を求め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

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

<table>
<tr><td colspan="6">原文</td></tr>
<tr><td colspan="6">With the improvements of technology, a new life with genetic risk can be detected.</td></tr>
<tr><td colspan="6">リファレンス</td></tr>
<tr><td colspan="6">With the improvements in technology, a new life with genetic risk can be detected.</td></tr>
<tr><td rowspan="3">訂正文 A</td><td colspan="5">With the improvement of technology, a new life with genetic risk can be detected.</td></tr>
<tr><td>人手評価</td><td>提案手法</td><td>M^2</td><td>I-measure</td><td>GLEU</td></tr>
<tr><td>5</td><td>0.884</td><td>0.0</td><td>-0.114</td><td>0.449</td></tr>
<tr><td rowspan="3">訂正文 B</td><td colspan="5">With the improvements in technology, a new life with genetic risk can be detected.</td></tr>
<tr><td>人手評価</td><td>提案手法</td><td>M^2</td><td>I-measure</td><td>GLEU</td></tr>
<tr><td>5</td><td>0.874</td><td>1.0</td><td>1.0</td><td>0.566</td></tr>
</table>

Table 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．

セットを用いて各システムに対する人手評価を TrueSkill [25] 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，人手評価が与えられている文のみを使用した．

また，全入力文に対する訂正を評価するために，参照有り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．GEC の先行研究と直接比較することができるように M^2 は，GEC システムの評価で最も一般的な方法に従い計算した．節 2.5.1，節 2.5.2 で用いた M^2 と異なるのは，参照文には公式の 2 セットのみを用いる点，システム単位のスコアが macro-$F_{0.5}$ 値によって算出される点である．GLEU については節 2.5.1，節 2.5.2 と同様，正確な評価のために参照文に 18 セット全てを用い，システム単位のスコアは文単位のスコアの平均によって算出した．

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．いずれの評価尺度でも参照無し手法で訂正を選択することにより訂正性能が向上した．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では無いが，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．

この実験結果から参照無し評価手法は，文法誤り訂正の性能向上に有用である

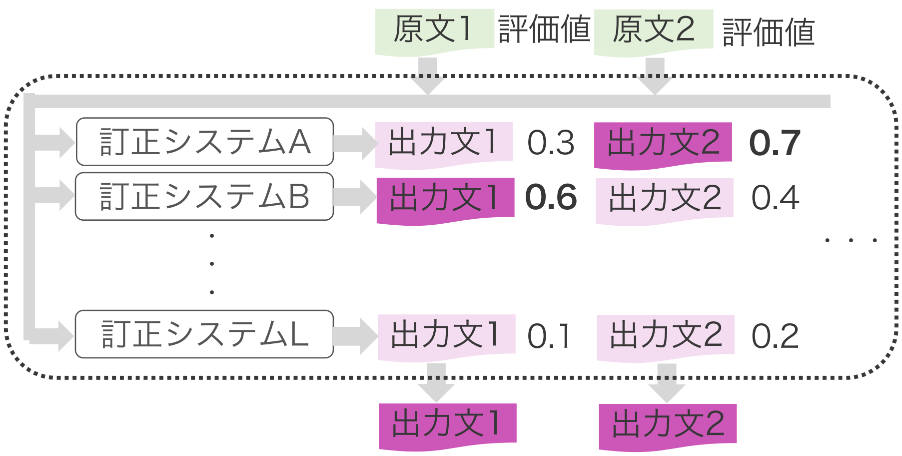


Figure 5: アンサンブルシステムの概要．各システムの訂正を参照無し手法によって評価し，最善の文を出力する．

評価尺度	アンサンブル	トップシステム
TrueSkill	0.462	0.191（AMU）
M^2	0.412	0.372（CAMB）
GLEU	0.548	0.531（CAMB）

Table 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し，括弧内にシステム名を示した．

と言える．また，本研究で行ったアンサンブル手法ではなく，参照無し評価手法のコンポーネントである文法性，流暢性，意味保存性の尺度を直接 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．アンサンブル手法は従来モデルの訂正候補から最良のものを選択するのに対し，そうしたモデルは３観点を考慮した訂正を出力できるため，さらなる性能向上が期待できる．

2.6　まとめ

本研究では，GEC システムを自動で評価するための参照無し手法を提案し，文法性，流暢性，意味保存性の観点を組み合わせることにより，GEC システムの自動評価を従来手法よりも正確に行える可能性があることを実験的に示した．また，文単位での評価性能を調べる実験を行ったところ，提案した参照無し手法が従来手法より高い性能を示した．さらに，参照無し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，参照無し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．

今後の展望としては，大量のデータを活用し，各観点の評価方法をより精緻な手法にすることで性能の向上を図ることが考えられる．例えば，誤り訂正の対訳コーパスから，ニューラルネットワークを用いて文法性を学習する手法が考えられる．また，3 観点の組み合わせ方を線形和ではなく，意味保存性のスコアが減点項として働くような組み合わせ方に変更することが考えられる．

3　ニューラル文法誤り訂正モデルによる反復訂正

3.1　背景

文法誤り訂正（Grammatical Error Correction）は，文法的な誤りを含む文から正しい文に訂正するタスクである．文法誤り訂正タスクでは機械翻訳の手法が用いられることが多い [28, 29, 30, 31]．機械翻訳では，原言語文と目的言語文の意味が等価であるという仮定がおけるため，意味的な情報を保持した言語の変換問題とみなすことができる．一方，文法誤り訂正では，機械翻訳のように原言語文と目的言語文の意味的な一致は同様に求められるが，それに加えて，文法的な正確性という追加の指標に合わせた言い換えを必要とする問題と捉えることができる．よって，文法誤り訂正では，機械翻訳よりも，個々の単語選択に関して文法的な整合性をより強く考慮することが求められる．このような観点から，文法誤り訂正では，段階的な訂正が必要な場合があると考えられる（図 6）．

この例のように，誤りの間に依存関係があるような場合，人間であれば適切な順序で訂正を行っていると考えられる．一方で機械翻訳モデルでは訂正文（出力）は，文頭から生成する処理形式となるため，文の末尾側にある誤りをどのように訂正するかという情報は，先頭側にある訂正の決定に影響を与えるのは困難である．しかし，文法誤り訂正は，機械翻訳とは違い，入力文と出力文の言語が同じであるという性質があるため，訂正文を再び入力として再度訂正させるという処理が可能である．よって，誤り箇所が多く，文の末尾側の誤りに起因して一回では全ての誤りを訂正するのが困難な文を，繰り返し処理により，徐々に改善できることが期待できる．

そこで，本研究では，段階的な訂正を適用した際の効果を検証する．実験では，文法誤り訂正で現在標準的に用いられている手法を用い，標準的に広く用いられ

The social network *plays* a role in providing information.

主語の数を訂正 ↓

Social networks ***plays*** a role in providing information.

動詞を主語と一致させる ↓

Social networks **play** a role in providing information.

Figure 6: 段階的な訂正が必要な例

ているベンチマークデータにて評価を行う．実験結果から，段階的な訂正を適用しても，期待通りの効果が得られないことを示す．また，段階的な訂正により，あまり効果が得られなかった原因を分析し，そこから導き出される知見を示す．

関連研究文法誤り訂正の分野は CoNLL-2014 shared task [24] が開催されて以来注目されている．近年の文法誤り訂正で高い性能を達成しているシステムはいずれもニューラル機械翻訳の手法が用いられている．Chollampatt ら [29] は Convolutional seq2seq [32] モデルが文法誤り訂正にも有効であることを示した．Junczys-Dowmunt ら [30] は Transformer [33] などのモデルを用いて low-resource MT で用いられている手法が文法誤り訂正にも有効であることを示した．

本研究のように訂正を繰り返すアプローチは昨今のいくつかの研究で行われている．Ge ら [31] は言語モデルに基づく流暢性スコアが訂正によって一定以上改善しない場合は，その出力を入力文として再学習を行うという fluency-boost learning を提案した．また，テスト時も流暢性スコアが改善しなくなるまで訂正するという fluency-boost inference を提案した．Lichtarge ら [34] はテスト時に，最も尤もらしい訂正システムの出力の尤度が，何も訂正されていない出力の尤度よりも一定以上改善しなくなるまで訂正を繰り返す，という手法を用いた．これら先行研究では，どちらも訂正の反復で性能が悪化することを防ぐ機構が導入されている．しかし，そうした機構を導入しない場合と比べたときの差分については調べられていない．つまり，反復訂正による実際の効果というのがあまり厳密には示されていないという問題点があると考えられる．

Table 9: 訂正の反復の結果

	CoNLL($F_{0.5}$)	JFLEG (GLEU)
無編集	0.0	40.54
1 回目	45.70	51.19
2 回目	46.11	51.79
3 回目	46.19	51.80
4 回目	46.24	51.81
5 回目	46.24	51.81

3.2 反復訂正

Ge ら [31] によると，seq2seq モデルは文法誤りを多く含む文を一度に完全に訂正することは通常できない．その理由として，文内文脈の読み取りが困難になる誤りが文中に含まれる場合，他の誤り箇所の訂正も困難になるということが挙げられる．一方で，文法誤り訂正は機械翻訳と異なり原言語と目的言語が同じであるため，文を同じモデルで何度も訂正することが可能である．そこで，まず原文を seq2seq モデルに入力し，そこで得られた出力を再び入力にする，という処理を繰り返すという方法論が考えられる．Ge ら [31] は流暢性スコアが改善しなくなるまで訂正を反復すると性能が向上することを示しているが，その性能向上は小さく，訂正を繰り返すことの実質的な効果や流暢性スコアを終了条件として用いる効果についての分析はなされていない．そこで，本研究では反復訂正によって実際に訂正の質がどのように変わるかについて詳細に調査を行う．

3.3 実験

前節で述べた通り，文法誤り訂正の標準的なモデルを用いて，訂正処理を繰り返し適用した場合に，性能がどのように変化するかを詳細に調査する．本実験では，Chollampatt ら [29] が用いた設定に従って，学習および評価をおこなった．

3.3.1 データセット

学習には公開されている Lang-8 Learner Corpora [35] と NUS Corpus of Learner English (NUCLE), [36] を用いた．Lang-8 に対しては Chollampatt ら [29] と同様に，英語話者によるエッセイと英語で書かれていないエッセイを除去する前処理

を行った．さらに，学習データから何も訂正されていない文対は除去した．また，NUCLE の約 10%に相当する約 5,400 文を開発データとして用いた．学習データには残りの約 130 万文を用いた．

3.3.2 性能評価

性能評価は文法誤り訂正のベンチマークとして用いられている設定で行った．すなわち，訂正システムの性能を CoNLL-2014 テストセット上において MaxMatch (M^2) scorer [7] で計算された $F_{0.5}$ 値によって評価した[8]．また，訂正システムの出力の流暢性を評価するために，JHU FLuency-Extended GUG corpus (JFLEG) [17] テストセット上において GLEU [9, 37] により評価した．

3.3.3 モデル設定

訂正モデルには CNN seq2seq [32] を用い，実装は Chollampatt ら [29] が Fairseq-py を拡張したものを用いた．この実装は現在文法誤り訂正の研究でよく使われているものの 1 つである [38, 39]．単語埋め込み層は 500 次元，ボキャブラリにターゲット側の頻度上位 30,000 の BPE ユニットを用いた．また単語埋め込み層の初期化は Chollampatt ら [29] が Wikipedia コーパス上で $fastText$ [40] を用いて学習したものを用いた．エンコーダおよびデコーダの層数はそれぞれ 7 とし，窓幅は 3，出力次元数は 1,024 次元に設定した．ドロップアウト率は 0.2 に設定し埋め込み層，畳み込み層，出力層に適用した．バッチサイズは 32 とし，最適化は Nesterov の加速勾配法を用い，学習率は 0.25, 慣性項は 0.99，クリップノルムは 0.1 とした．各エポック終了後にバリデーションを行い，ロス関数の値が最良のモデルをテストに用いた．デコードは窓幅 12 のビーム探索を用いた．出力の 1 ベストを次ステップでは入力として用いた．

3.3.4 結果

反復訂正による CoNLL-2014 テストデータにおける M^2 スコアと JFLEG テストデータにおける GLEU スコアを表 9 に示す．

訂正を繰り返すことによって性能がわずかに（$F_{0.5}$ で 0.5 ポイント）改善した．これはテスト時に流暢性のスコアが上昇しなくなるまで訂正するという方法を用いた先行研究 [31] と同程度の改善幅である．本実験では 3 回程度の訂正で性能が

[8]テストデータ 1,312 文中 3 文については単語数が多く評価に時間がかかるため除外した

収束した．また，訂正後の出力を文単位で比較したところ，CoNLL-2014のテストデータ1,309文中，5回の訂正で1回目より改善したのは16文，悪化したのは13文であった．JFLEGでは747文中5回目の訂正で1回目より改善した文数は50文，悪化したのは20文であった．このことから，訂正を反復してもほとんどの出力は変わらないといえる．原因としては，訂正モデルが系列全体を一度に訂正するという学習は行っているが，文法誤り訂正の出力結果で訂正しきれていない誤り、つまり訂正後の文に残っている誤り (システムで訂正が難しい事例) に対する学習データが存在していないため，それらを訂正するような学習をしていないことが考えられる．この仮説が正しいとすると，現在の用いている学習データに対して，複数の誤りがあるデータを正解に向けて一部修正したデータを追加の学習データとして活用する方法などが考えられる．

また，段階的な訂正が必要な事例が少ないという可能性も考えられる．段階的に訂正が必要な誤りには少なくとも3種類が考えられる: (1) 1つの単語に複数の誤りが複合的に入っている場合（表10のobvioualyのように綴り誤りと文法誤りが混じっている場合），(2) 前方の誤った単語を訂正しないと後方の誤りを訂正できない場合，(3) 後方の誤りを訂正しないと前方の誤りを訂正できない場合がある．このうち2番目に関しては，従来の文法誤り訂正システムは前から訂正するため，システムが正しく訂正できていれば後半の誤りに関しても訂正できる可能性がある。一方，1番目と3番目に関しては，一度に訂正することは難しいがこのような例がテストデータ中には出現していない可能性が考えられる．

この仮説を検証するため，CoNLL 1,309文の2つの正解データを調査した．1,309文中，文内での訂正箇所が両アノテータとも1箇所以下であるものが397文，どちらかが2箇所に誤りがあるとしたものが461文，両者とも2箇所に誤りがあるとしたものは53文あり，今回はこの53文を本論文著者が分析した．この53文中，両者のアノテータが上記で述べた (1) のように複合した誤りを訂正している文は1文，(2)，(3) のようにある誤りが他の箇所の誤りに影響している文は2文しか無く，アノテータの訂正が両者とも独立な箇所の誤りを訂正していると判断されたものは45文であった．このことから段階的な訂正が必要な事例は少ないことが考えられる．一方で，この53文のうち，訂正システムが1度の訂正で2箇所とも直せたのは2文，また2回目以降の訂正で出力が変化したのは2文であり，訂正モデルにとっては誤り箇所が減っても訂正は容易になっていないことが観測された．これは，実験で用いたモデルの性能のPrecisionが60.07，Recallが23.36と，Precisonを重視していることも一因と考えられる．

3.3.5 事例分析

訂正を反復することで出力が改善した例 3 つを表 10 に示す．例 (a) には 2 箇所の文法誤りと 1 箇所の綴り誤りがあり，訂正モデルは 1 度目の訂正で 2 つの文法誤りに関しては正しく訂正できたが，綴り誤りに関しては訂正できず，さらに綴り誤りの影響を受けて “litterature” の前に不要な the を挿入した，しかし 2 回目の訂正で “litterature” の綴り誤りを修正し，3 回目の訂正で “the” を削除して出力を改善できている．例 (b) では，綴り誤りと語形誤りが複合した誤りを段階的に訂正できた例で，1 回目の訂正で綴り誤りを訂正したことで 2 回目で副詞の “obviously” から形容詞 “ovbious” に訂正できた．例 (c) では，1 回目の訂正で後半の “thy” を “they” に直せたため，2 回目で前半にある “Thy” や “there self” を “They” と “themselves” にそれぞれ訂正することに成功している．

一方，訂正を繰り返すことで性能が悪くなった例を表 11 に示す，例 (d)1 回目は適切な訂正であるが，2 回目の訂正で the を誤って挿入したため $F_{0.5}$ スコアが低下した．例 (e) の 2 回目の訂正のように，文をさらに流暢にしようとするあまり正誤の判定が難しい訂正を行っているような例も見られた．

4 終わりに

本研究では，人間らしい文法誤り訂正のために，まず参照無し自動評価手法を提案した．提案手法はベンチマークデータ上において従来の参照有り評価よりも正確な評価を行うことができた．また，提案手法は訂正システムの性能の向上に役立つ可能性を示した．次に，文法誤り訂正において訂正を繰り返すことによる効果を調査した．文法誤り訂正では誤りを多く含むような文を一度に全て正しく訂正するのは困難な場合があり，そのような文が繰り返し処理により改善されていくことが期待されたが，実験の結果，効果は限定的なものであり，多くの文では 2 回目以降の訂正が行わないことがわかった．一方で，訂正が悪化する事例よりは改善する事例の方が多く見られた．また，事例分析を通して訂正システムの今後の課題を考察した．

謝辞

本研究を進めるにあたり，多くの皆様のご協力，ご助言をいただきましたことに，ここに心より感謝申し上げます．主指導教員である乾健太郎教授には，ご多忙の

中，研究活動だけでなく進路に関することなど多くのご指導，ご助言を頂きましたことに心より感謝申し上げます．副指導教員である鈴木潤准教授には，同じく研究活動に関して多くのご助言を頂きましたことに心より感謝申し上げます．ご多忙の中審査委員をお引き受けくださいました，張山昌論教授，篠原歩教授に心より感謝申し上げます．研究方針や研究手法，論文執筆に関しまして，直接のご指導を頂いた水本智也特任研究員に心より感謝申し上げます．研究会などで多くのご助言を頂いた松林優一郎研究員に心より感謝申し上げます．また，日々の議論におきまして，多くのアドバイスを頂きました乾・鈴木研究室の皆様に感謝申し上げます．最後になりましたが，学校生活におきまして関わってくださいましたすべての皆様に感謝致します．

References

[1] 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. Association for Computational Linguistics, 2016.

[2] 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.

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72. Association for Computational Linguistics, 2005.

[4] Ondřej Bojar, Yvette Graham, and Amir Kamran. Results of the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pp. 489–513. Association for Computational Linguistics, 2017.

[5] 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. Association for Computational Linguistics, 2015.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[7] 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, 2012.

[8] Mariano Felice and Ted Briscoe. Towards a standard evaluation method for grammatical error detection and correction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 578–587. Association for Computational Linguistics, 2015.

[9] 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. Association for Computational Linguistics, 2015.

[10] Robert Dale and Adam Kilgarriff. Helping our own: The hoo 2011 pilot shared task. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pp. 242–249. Association for Computational Linguistics, 2011.

[11] Robert Dale, Ilya Anisimoff, and George Narroway. Hoo 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pp. 54–62. Association for Computational Linguistics, 2012.

[12] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pp. 131–198. Association for Computational Linguistics, 2016.

[13] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 conference on

machine translation. In *Proceedings of the Second Conference on Machine Translation*, pp. 169–214. Association for Computational Linguistics, 2017.

[14] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pp. 223–231, 2006.

[15] Marcin Miłkowski. Developing an Open-source, Rule-based Proofreading Tool. *Software: Practice and Experience*, Vol. 40, No. 7, pp. 543–566, 2010.

[16] Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 174–180. Association for Computational Linguistics, 2014.

[17] 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. Association for Computational Linguistics, 2017.

[18] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. *English Gigaword Fifth Edition LDC2011T07*. Philadelphia: Linguistic Data Consortium, 2011.

[19] Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. *TOEFL11: A Corpus of Non-Native English*. Technical report, Educational Testing Service, 2013.

[20] Jey Han Lau, Alexander Clark, and Shalom Lappin. Unsupervised prediction of acceptability judgements. In *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. Association for Computational Linguistics, 2015.

[21] Tomáš Mikolov. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 2012.

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

[23] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380. Association for Computational Linguistics, 2014.

[24] 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. Association for Computational Linguistics, 2014.

[25] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: A bayesian skill rating system. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pp. 569–576. MIT Press, 2007.

[26] Christopher Bryant and Hwee Tou Ng. How far are we from fully automatic high quality grammatical error correction? In *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. Association for Computational Linguistics, 2015.

[27] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 521–535, 2016.

[28] Roman Grundkiewicz and Marcin Junczys-Dowmunt. Near human-level performance in grammatical error correction with hybrid machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 284–290. Association for Computational Linguistics, 2018.

[29] Shamil Chollampatt and Hwee Tou Ng. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 5755–5762, 2018.

[30] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 595–606. Association for Computational Linguistics, 2018.

[31] Tao Ge, Furu Wei, and Ming Zhou. Fluency boost learning and inference for neural grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1055–1065. Association for Computational Linguistics, 2018.

[32] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. *CoRR*, Vol. abs/1705.03122, , 2017.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NIPS*, pp. 5998–6008. Curran Associates, Inc., 2017.

[34] Jared Lichtarge, Christopher Alberti, Shankar Kumar, Noam Shazeer, and Niki Parmar. Weakly supervised grammatical error correction using iterative decoding. *CoRR*, Vol. abs/1811.01710, , 2018.

[35] Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. Mining revision log of language learning sns for automated japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 147–155. Asian Federation of Natural Language Processing, 2011.

[36] Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*.

[37] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel R. Tetreault. GLEU without tuning. *CoRR*, Vol. abs/1605.02592, , 2016.

[38] Shamil Chollampatt and Hwee Tou Ng. Neural quality estimation of grammatical error correction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2528–2539. Association for Computational Linguistics, 2018.

[39] Tao Ge, Furu Wei, and Ming Zhou. Reaching human-level performance in automatic grammatical error correction: An empirical study. *CoRR*, Vol. abs/1807.01270, , 2018.

[40] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017.

発表文献一覧

学術論文誌

- 浅野広樹, 水本智也, 乾健太郎. 文法性・流暢性・意味保存性に基づく文法誤り訂正の参照無し評価. 自然言語処理, Vol.25, No.5, pp.555-576, December 2018.

国際会議論文

- 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 8th International Joint Conference on Natural Language Processing (IJCNLP), pp.343-348, November 2017.

国内会議・研究会論文

- 浅野広樹, 水本智也, 松林優一郎, 乾 健太郎. 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. 第 4 回自然言語処理シンポジウム・第 234 回自然言語処理研究会, 8 pages, December 2017.

- 浅野広樹, 水本智也, 乾健太郎. 文法誤り訂正のためのリファレンスレス評価. 言語処理学会第 23 回年次大会, pp.947-950, March 2017.

- 浅野広樹, 水本智也, 乾健太郎. 文法誤り訂正における反復訂正の効果検証. 言語処理学会第 25 回年次大会，March 2019 (to appear).

Table 10: スコアが改善された例

	訂正回数	文
(a)	0	People *tends* to choose other *medias* , and that is why *litterature* is in danger .
	1	People **tend** to choose other **media** , and that is why *the litterature* is in danger .
	2	People tend to choose other media , and that is why *the* **literature** is in danger .
	3	People tend to choose other media , and that is why ϕ literature is in danger .
(b)	0	On one side , it is *obviousaly* that many advantages have been brought to our lives .
	1	On one side , it is **obviously** that many advantages have been brought to our lives .
	2	On one side , it is **obvious** that many advantages have been brought to our lives .
(c)	0	*Thy* are busy *in there self* , *thy dont* spend time to help the society that they live in .
	1	*Thy* are busy *in there self* , **they do n't** spend time to help the society that they live in .
	2	**They** are busy **themselves** , they do n't spend time to help the society that they live in .

Table 11: スコアが悪化した例

	訂正回数	文
(d)	0	And we keep track of all family *members* health conditions .
	1	And we keep track of all family **members '** health conditions .
	2	And we keep track of all **the** family members ' health conditions .
(e)	0	With the willing of people to become the best among others , social network *site* now become a place to show off .
	1	With the willing of people to become the best among others , social network **sites** now become a place to show off .
	2	With the **willingness** of people to become the best among others , social **networking** sites now become a place to show off .