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

浅野 広樹^{†,††}・水本 智也^{††}・乾 健太郎^{†,††}

文法誤り訂正の研究開発では、訂正システムの性能を自動評価することは重要であると考えられている。従来の自動評価手法では参照文が必要であるが、参照文は人手で作成しなければならないため、コストが高く網羅性に限界がある。この問題に対処するために、参照文を用いず、文法性の観点によって訂正を評価する参照無し手法が提案されたが、従来の参照有り手法の性能を上回ることはできなかった。そこで本研究では、先行研究で提案された手法を拡張し、参照無し手法の可能性について調査する。具体的には、文法性に加えて流暢性と意味保存性を組み合わせた参照無し手法が、従来の参照有り手法よりも人手評価スコアを正確に予測できることを実験的に示す。また、参照無し手法は文単位でも適切な評価が可能であることと、文法誤り訂正システムに応用可能であることを示す。

キーワード：文法誤り訂正，自動評価

A Reference-less Evaluation Metric Based on Grammaticality, Fluency, and Meaning Preservation in Grammatical Error Correction

HIROKI ASANO^{†,††}, TOMOYA MIZUMOTO^{††} and KENTARO INUI^{†,††}

In grammatical error correction (GEC), the automatic evaluation of system performance is thought to be an essential driving force. Previous methods for automated system assessment require gold-standard references, which have to be created manually and thus tend to be both expensive and limited in coverage. To address this problem, a reference-less approach has recently emerged; however, previous reference-less metrics, which only consider the grammaticality of system outputs, have not performed as well as reference-based metrics. In this study, we explore the potential of extending a prior grammaticality-based method to establish a reference-less evaluation method for GEC systems. We empirically show that a reference-less metric that combines both fluency and meaning preservation with grammaticality provides a better estimate of manual scores than that of commonly used reference-based metrics. Additionally, we show that the reference-less metric can provide appropriate evaluation at the sentence-level and that it can be applied to GEC systems.

[†] 東北大学情報科学研究科, Graduate School of Information Sciences, Tohoku University

^{††} 理化学研究所革新知能統合研究センター, RIKEN Center for Advanced Intelligence Project

Key Words: *Grammatical Error Correction, Evaluation Metric*

1 序論

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。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、Napoles, Sakaguchi, and Tetreault (2016) 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 (Sakaguchi, Napoles, Post, and Tetreault 2016) 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 Napoles et al. (2016) 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、Napoles et al. (2016) 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ろと考えられる。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 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 (Banerjee and Lavie 2005). 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 (Bojar, Graham, and Kamran 2017) 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 (Grundkiewicz, Junczys-Dowmunt, and Gillian 2015; Sakaguchi et al. 2016; Napoles et al. 2016). 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 5.1.1 節、5.2.2 節でそれぞれ詳述する.

3 既存の評価尺度

機械翻訳の分野では、BLEU (Papineni, Roukos, Ward, and Zhu 2002) 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.

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 (Dahlmeier and Ng 2012), I-measure (Felice and Briscoe 2015), GLEU (Napoles, Sakaguchi, Post, and Tetreault 2015; Sakaguchi et al. 2016) が

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.

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 (Dale and Kilgarrieff 2011; Dale, Anisimoff, and Narroway 2012)。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 (Papineni et al. 2002) 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。

$$\text{GLEU+} = \text{BP} \cdot \exp \left(\sum_{n=1}^4 \frac{1}{n} \log(p_n') \right) \quad (1)$$

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

ただし、 $N(A, B, C, \dots)$ 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 (Napoletano et al. 2016)。

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 (Bojar, Chatterjee, Federmann, Graham, Haddow, Huck, Jimeno Yepes, Koehn, Logacheva, Monz, Negri, Neveol, Neves, Popel, Post, Rubino, Scarton, Specia, Turchi, Verspoor, and Zampieri 2016; Bojar, Chatterjee, Federmann, Graham, Haddow, Huang, Huck, Koehn, Liu, Logacheva, Monz, Negri,

Post, Rubino, Specia, and Turchi 2017). 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-targeted Translation Error Rate (HTER) (Snover, Dorr, Schwartz, Micciulla, and Makhoul 2006) 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 Napoles et al. (2016) 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でないため, 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 Napoles et al. (2016) 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 (Miłkowski 2010) による文法誤り検出数に基づく評価, 3 つ目は Heilman, Cahill, Madnani, Lopez, Mulholland, and Tetreault (2014) 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 Heilman et al. (2014) のモデルなどを組み合わせることで性能向上を図る.

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, Napoles et al. (2016) 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 つの観点から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. Napoles, Sakaguchi, and Tetreault (2017) 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 (Napoles et al. 2016). 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 (Sakaguchi et al. 2016) において文

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.

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。

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

ただし S_G, S_F, 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 $\text{Score}(h, s)$ 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。

4.1 文法性

Napoles et al. (2016) が参照無し評価に用いたモデルのうち、Heilman et al. (2014) 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 $S_G(h)$ 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、Heilman et al. (2014) 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 Tool³による誤り検出数を素性として用いた。

モデルの学習は GUG データセット (Heilman et al. 2014) に対して Napoles et al. (2016) の実装⁴を用いた。さらに、言語モデルの学習のために Gigaword (Parker, Graff, Kong, Chen, and Maeda 2011) と TOEFL11 (Blanchard, Tetreault, Higgins, Cahill, and Chodorow 2013) 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 Napoles et al. (2016) 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。

4.2 流暢性

文法誤り訂正における流暢性の重要性は (Sakaguchi et al. 2016; Napoles et al. 2017) 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 (Lau, Clark, and Lappin 2015)。具体的には、訂正文 h に対し、流暢性 $S_F(h)$ を次のように求める⁵。

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

³ <https://languagetool.org>

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

$|h|$ 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 (Mikolov 2012) を採用し, 実装は faster-rnnlm⁶を用いた. 学習には British National Corpus (BNC Consortium 2007) および Wikipedia の 1,000 万文を用いた. 作成したモデルは Lau et al. (2015) 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.

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評価尺度を用いる方法が考えられる. 本研究では, METEOR (Denkowski and Lavie 2014) 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 $S_M(h, s)$ を次式により求める.

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

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

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

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.

⁶ <https://github.com/yandex/faster-rnnlm>

5 実験

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。

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。

5.1.1 実験設定

Napoles et al. (2016) 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 (Grundkiewicz et al. 2015) 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 (Ng, Wu, Briscoe, Hadiwinoto, Susanto, and Bryant 2014) のデータセット、およびそれに対して Grundkiewicz et al. (2015) 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 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、Grundkiewicz et al. (2015) 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 (Herbrich, Minka, and Graepel 2007) を用いて訂正システム単位の手人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、Bryant and Ng (2015) による参照文が 8 個、Sakaguchi et al. (2016) 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

図 1 自動評価尺度のシステム単位評価

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 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

図 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。

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.

5.1.2 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

図3 図2左をz軸方向から見た図

表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

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

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、文法性は α, β, γ 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 (Linzen, Dupoux, and Goldberg 2016)、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。

5.2 文単位評価の性能調査

5.1 節の実験で、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ます。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の程

度正確に評価できるかを調査する.

5.2.1 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 Grundkiewicz et al. (2015) 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手人評価のために作られたものではあるが, 訂正システムの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, 1 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.

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.

図 4 文単位評価が不適切な例

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

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 <i>higher</i> for the treatment.	5
h_3 : Genetic diseases cost <i>high</i> for the treatment.	3
h_4 : Genetic diseases <i>costs</i> highly for treatment.	3

5 が最も良く, 1 が最も悪い.

5.2.2 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るにもかかわらず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。

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

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。

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

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。

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

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 Grundkiewicz et al. (2015) 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。

5.2.3 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ます. 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.

5.2.4 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

表 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	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

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

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

照無し手法は A の方を高く評価できたが, 参照有り手法は 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 2 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

表 5 リファレンスペース手法の優劣判定の誤り例

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.				
	人手評価	提案手法	M ²	I-measure	GLEU
	3	0.892	0.00	-0.391	0.414
訂正文 B	On the other hand, viewer <i>are</i> not listeners.				
	人手評価	提案手法	M ²	I-measure	GLEU
	2	0.651	0.714	-0.096	0.496

人手評価は 5 が最も良く, 1 が最も悪い.

表 6 リファレンスレス手法の優劣判定の誤り例

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, <i>the</i> carrier of ϕ known genetic risk should not be obligated to tell his or her relatives.				
	人手評価	提案手法	M ²	I-measure	GLEU
	4	0.857	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.				
	人手評価	提案手法	M ²	I-measure	GLEU
	5	0.851	0.625	0.222	0.348

人手評価は 5 が最も良く, 1 が最も悪い.

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

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 <i>in</i> technology, a new life with genetic risk can be detected.					
訂正文 A	With the <i>improvement</i> of technology, a new life with genetic risk can be detected.				
	人手評価	提案手法	M ²	I-measure	GLEU
	5	0.884	0.0	-0.114	0.449
訂正文 B	With the improvements <i>in</i> technology, a new life with genetic risk can be detected.				
	人手評価	提案手法	M ²	I-measure	GLEU
	5	0.874	1.0	1.0	0.566

自動評価スコアは標準化前の値.

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 M² 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.

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

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のの可能性を調査する. 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.

5.3.1 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.

5.3.2 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, Grundkiewicz et al. (2015) や Naples et al. (2017) 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 らのデータセットを用いて各システムに対する人手評価を TrueSkill (Herbrich et al. 2007) 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 (1,312 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. 5.1 節, 5.2 節で用いた 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 5.1 節, 5.2 節と同様, 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.

図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.

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

表 8 訂正システムに対するスコア

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)

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。

5.3.3 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ないが、明らかな性能向上を示している。M² 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と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。

6 結論

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。

謝 辞

本論文の査読にあたり, 著者の不十分な記述などに対してご意見・ご指摘をくださった査読者の方々へ感謝します. 本論文の内容の一部は, 情報処理学会第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(浅野, 水本, 松林, 乾 2017) および The 8th International Joint Conference on Natural Language Processing (Asano, Mizumoto, and Inui 2017) で発表されたものです.

参考文献

- Asano, H., Mizumoto, T., and Inui, K. (2017). “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 (Volume 2: Short Papers)*, pp. 343–348. Asian Federation of Natural Language Processing.
- 浅野広樹, 水本智也, 松林優一郎, 乾健太郎 (2017). 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. 情報処理学会第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.
- Banerjee, S. and Lavie, A. (2005). “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.
- Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. (2013). *TOEFL11: A Corpus of Non-Native English*. Technical report, Educational Testing Service.
- BNC Consortium (2007). *The British National Corpus*. version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.
- Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). “Findings of the 2017 Conference on Machine Translation.” In *Proceedings of the 2nd Conference on Machine Translation*, pp. 169–214. Association for Computational Linguistics.
- 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. (2016). “Findings of the 2016 Conference on Machine Translation.” In *Proceedings of*

- the 1st Conference on Machine Translation*, pp. 131–198. Association for Computational Linguistics.
- Bojar, O., Graham, Y., and Kamran, A. (2017). “Results of the WMT17 Metrics Shared Task.” In *Proceedings of the 2nd Conference on Machine Translation*, pp. 489–513. Association for Computational Linguistics.
- Bryant, C. and Ng, H. T. (2015). “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.
- Dahlmeier, D. and Ng, H. T. (2012). “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.
- Dale, R., Anisimoff, I., and Narroway, G. (2012). “HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task.” In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP*, pp. 54–62. Association for Computational Linguistics.
- Dale, R. and Kilgariff, A. (2011). “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.
- Denkowski, M. and Lavie, A. (2014). “Meteor Universal: Language Specific Translation Evaluation for Any Target Language.” In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pp. 376–380. Association for Computational Linguistics.
- Felice, M. and Briscoe, T. (2015). “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.
- Grundkiewicz, R., Junczys-Dowmunt, M., and Gillian, E. (2015). “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.
- Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M., and Tetreault, J. (2014). “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.

- Herbrich, R., Minka, T., and Graepel, T. (2007). “TrueSkillTM: A Bayesian Skill Rating System.” In Schölkopf, B., Platt, J. C., and Hoffman, T. (Eds.), *Advances in Neural Information Processing Systems 19*, pp. 569–576. MIT Press.
- Lau, J. H., Clark, A., and Lappin, S. (2015). “Unsupervised Prediction of Acceptability Judgments.” 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.
- Linzen, T., Dupoux, E., and Goldberg, Y. (2016). “Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies.” *Transactions of the Association for Computational Linguistics*, **4**, pp. 521–535.
- Mikolov, T. (2012). “Statistical Language Models Based on Neural Networks.” *Presentation at Google, Mountain View, 2nd April*.
- Miłkowski, M. (2010). “Developing an Open-source, Rule-based Proofreading Tool.” *Software: Practice and Experience*, **40** (7), pp. 543–566.
- Napoles, C., Sakaguchi, K., Post, M., and Tetreault, J. (2015). “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.
- Napoles, C., Sakaguchi, K., and Tetreault, J. (2016). “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.
- Napoles, C., Sakaguchi, K., and Tetreault, J. (2017). “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.
- Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (2014). “The CoNLL-2014 Shared Task on Grammatical Error Correction.” In *Proceedings of the 18th Conference on Computational Natural Language Learning: Shared Task*, pp. 1–14. Association for Computational Linguistics.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). “BLEU: a Method for Automatic Evaluation of Machine Translation.” In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318.

- Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). *English Gigaword 5th Edition LDC2011T07*. Philadelphia: Linguistic Data Consortium.
- Sakaguchi, K., Napoles, C., Post, M., and Tetreault, J. (2016). “Reassessing the Goals of Grammatical Error Correction: Fluency Instead of Grammaticality.” *Transactions of the Association for Computational Linguistics*, 4, pp. 169–182.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). “A study of translation edit rate with targeted human annotation.” In *Proceedings of Association for Machine Translation in the Americas*, pp. 223–231.

略歴

浅野 広樹：1993 年生。2017 年東北大学工学部情報科学研究科卒業。同年，同大学大学院情報科学研究科修士課程システム情報科学専攻に進学。現在に至る。自然言語処理の研究に従事。

水本 智也：2015 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。博士（工学）。同年より東北大学情報科学研究科特任助教を経て 2017 年より理化学研究所革新知能統合研究センター特別研究員。文法誤り訂正や記述式答案の自動採点など自然言語処理による言語学習／教育支援に関心がある。情報処理学会，人工知能学会，言語処理学会，ACL 各会員。

乾 健太郎：東北大学大学院情報科学研究科教授。1995 年東京工業大学大学院情報理工学研究科博士課程修了。同大学助手，九州工業大学助教授，奈良先端科学技術大学院大学助教授を経て，2010 年より現職。2016 年より理化学研究所 AIP センター自然言語理解チームリーダー兼任。情報処理学会論文誌編集委員長，同会自然言語処理研究会主査等を歴任，2018 年より言語処理学会論文誌編集委員長。

(2018 年 5 月 1 日 受付)

(2018 年 8 月 1 日 再受付)

(2018 年 9 月 12 日 採録)