B3TB2009

卒業論文

文法誤り訂正システムのリファレンスレス評価に関する研究

浅野 広樹

2017 年 3 月 31 日

東北大学
工学部 情報知能システム総合学科

文法誤り訂正システムのリファレンスレス評価に関する研究*

浅野 広樹

内容梗概

従来の GEC システムの評価ではリファレンスを用いた評価尺度が用いられてきた．しかしリファレンスありの評価では人手との相関が高い評価を実現するためには多くのリファレンスが必要であり，リファレンスの作成コストが高い．この問題を解決するために，文意の保存，文法性，自然さの観点を考慮したリファレンスレス評価を提案した．提案手法はリファレンスを用いないにも関わらず従来手法よりも人手との相関が高い評価を実現した．また，母国語話者による流暢な英文に対しても適切にスコアリングできていることを示した．

キーワード

文法誤り訂正，自動評価，言語モデル，ロジスティック回帰

*東北大学 工学部 情報知能システム総合学科 卒業論文, B3TB2009, 2017 年 3 月 31 日.

A Study of Reference-less Evaluation in Grammatical Error Correction*

Hiroki Asano

Abstract

Current methods for automatically evaluating grammmatical error correction systems rely on gold-standard references. However, these methods need a lot of gold-standard references in order to evaluate error correction systems like human, and it is costly to make many references. To solve this problem, we propose a reference-less metric which evaluates meaning preservation, grammaticality, and fluency. We achieve stronger correlation with human judgements than previous methods. Moreover, we show that our metric can appropriately evaluate fluent sentences written by native speakers.

Keywords:

Grammatical Error Correction (GEC), Automatic Evaluation, Language Model, Logistic Regression

*Graduation Thesis, Department of Information and Intelligent Systems, Tohoku University, B3TB2009, March 31, 2017.

目 次

図 目 次

表 目 次

1　はじめに

文法誤り訂正（Grammatical Error Correction: GEC）は，英語学習支援を目的とし，学習者の英文を自動で訂正するタスクである．近年 GEC の研究が注目されており，英語の文法誤りの訂正の性能を競うコンペティションが 2011 年から 4 年連続で開催された．GEC システムの研究開発においてはシステムの出力を人手で評価するのは大変であるため，自動評価が必要とされている．

GEC システムの自動評価は，正解データ（リファレンス）を用いる評価尺度が伝統的に使われている．リファレンス有り評価は文法誤りが訂正された数を数えることによって文法性の観点や，訂正の前後で文意が保存されているかという観点（文意の保存）を評価している．しかし，問題点として，訂正後の文が原文よりも良い文であるかどうかがスコアに反映されない点，リファレンスには無いが妥当な訂正をしたシステムが過小評価される点，リファレンス作成者が誤りを見逃した場合に同様の誤りを訂正しなかったシステムが過大評価される点が指摘されている [1][2]．この問題を解決するために多くのリファレンスを作成することがあるが，リファレンスの作成はコストが高く，妥当な訂正を網羅することは現実的ではない．

Napoles ら [2] はこうした問題点に対処するために，リファレンスを用いずに評価する手法（**リファレンスレス評価**）を提案した．彼らは文法性の観点のみで GEC システムを評価した．しかし人手評価との相関がリファレンスを用いる手法を超えることはできなかった．問題点としては，従来のリファレンスを用いる手法では考慮されていた文意の保存が考慮されていない点が挙げられる．また，Sakaguchi ら [3] は同じ文法的な文でも，より自然な文の方が人間に好まれることを確かめ，GEC システムの評価には英文の自然さを考慮する必要があることを示唆したが，Napoles らの手法ではこの観点を欠いている．

そこで，本研究では GEC のための新たなリファレンスレス評価手法を提案する．従来のリファレンスレス評価で考慮されていた文法性に加え，文意の保存と自然さを考慮することで GEC システムを評価する．実験の結果，文法性，文意の保存，自然さの 3 つを考慮した提案手法の評価性能は従来手法を上回った．また，自然さを考慮して GEC システムの評価を行うのは我々が初めてであり，GEC の

評価において自然さが重要であることを実験的に確認した.

2　関連研究

従来，GEC システムは CoNLL2014 Shared Task [4] のテストデータを用いて M^2 によって評価されてきた．M^2 はシステムの訂正とリファレンスを比較して $F_{0.5}$ 値を算出する方法である．Napoles ら [5] は出力文と正解文の n グラム一致率から，原文と出力文には存在し正解文に存在しない n グラムの存在率を減点することによって評価する GLEU を提案した．一方，リファレンスの網羅性を高めることも行われてきた [6]．Sakaguchi ら [3] はできるだけ原文を変えない「最小の修正」と訂正数を気にせず母国語話者のような文を目指す「流暢な修正」を CoNLL 2014 のテストデータに対して付与した．

Grundkiewicz ら [7] は GEC システムの自動評価と人手評価との相関を検証した．彼らはその検証のために CoNLL2014 Shared Task の参加チームの出力に対してスコアをつけた．そのスコアは各文に対する各システムの出力を人手でランキングし，それを元に算出した．

Napoles らは機械翻訳の評価と同様に文法性と文意の保存の観点でシステムを評価する手法を提案した．文法性はリファレンスを用いずに評価し，文意の保存はリファレンスを用いた M^2，GLEU によって評価し，これらの値を線形補間してスコアとした．各手法による評価性能は Grundkiewicz らの人手評価との相関係数で比較した．彼らはリファレンスを用いる手法とリファレンスレス評価を組み合わせる手法が従来手法よりも優れていること，コーパスのスコアは文単位のスコアの平均によって求める方が評価性能が良いことを示した．

3　提案手法

3.1　アイディア

GEC システムの自動評価には文法性，自然さ，文意の保存の 3 つの観点が重要である．文法性は，訂正後の文に文法的な誤りがあるかどうかを評価する．自然さは，GEC システムの出力がどのくらい自然な英文であるかを評価する．文意の保存は，訂正の前後で文意が変わっていないかを評価する．

自然さは母国語話者による英語の評価に影響する．Sakaguchi らは学習者の文に「最小の修正」と「流暢な修正」を付与し，流暢な修正の方が人手評価が高いことを示した．次の例における (1a) が原文，(1b) が最小の修正，(1c) が流暢な修正の 1 例である．

(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 .

(1b) と (1c) はどちらも文法的であるが，(1c) のような流暢な修正を高く評価するために自然さの観点が必要である．

文意の保存は，原文の意味が変わる訂正にペナルティを与えるためのものである．GEC では学習者の書いた文の意味を極力変えるべきではないという指針がある．例えば次のような非文 (2a) は，(2b) のように文法的でない箇所を消去すれば文法性の評価は上がるが，訂正としては不適切である．

(2) a. It is unfair to release a law only point to the genetic disorder.

b. It is unfair to pass a law.

このように，訂正を適切に評価するためには，文法性，自然さ，文意の保存を総合した評価が必要である．

3.2 モデル

3.1 節で述べた 3 つの観点を組み合わせた評価手法について述べる．次式のように，各観点によるスコアの重み付き和を訂正の良さを表すスコアとする．

$$Score = \alpha S_G + \beta S_N + \gamma S_M \ (\alpha + \beta + \gamma = 1) \tag{1}$$

文法性のスコア S_G，自然さのスコア S_N，文意の保存のスコア S_M の重み付き和を文のスコアとする．単純な線形和を採用したのは，学習者が文法性と自然さのどちらを重視するかで重みを任意に設定できるようにするためである．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．

文法性　Napoles らは文法誤り検出器を用いる手法と，Heilman ら [8] の言語学的な素性を用いる教師あり学習による手法を用いて文法性を評価した．本研究でも文法性については同様の手法で評価する．検出器による手法は，検出された誤りの数が多いと文法性は低くなるという仮定に基づき，文法性を次式で算出する．

$$S_G = 1 - \frac{\text{誤り数}}{\text{単語数}} \tag{2}$$

本研究では検出器として LanguageTool 3.5 を用いた．

Heilman らの手法はスペルミス数，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を用いて，ロジスティック回帰により与えられた文が文法的である確率を算出して S_G とする．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．

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．本研究では Lau ら [9] と同様に，自然さを次式で算出する [2]．

$$S_N = \frac{\log P_m(\xi) - \log P_n(\xi)}{|\xi|} \tag{3}$$

[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

ξ は文 ($|\xi|$ 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 と Wikipedia の合計 1000 万文で訓練した．出現頻度が低い単語は UNK に置換し，さらに一部の単語は表層的な特徴を付与したトークンに置換した（例: 1945 → UNK-NUM）．

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．本研究では訂正前後の文に METEOR 1.5 [10] 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．

$$P = \frac{m(h_c)}{|h_c|} \tag{4}$$

$$R = \frac{m(r_c)}{|r_c|} \tag{5}$$

$$S_M = \frac{P \cdot R}{t \cdot P + (1-t) \cdot R} \tag{6}$$

h_c は GEC システムの出力中の内容語，r_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，$m(r_c)$ は原文中の内容語でアライメントされた単語数を表す．原文中の内容語は冗長性などの理由により削除されることがあるため，$t = 0.85$ を用いた．

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

4　実験

3 節で提案した手法の性能を調べるために，GEC システムの出力を用いて提案手法によるスコアと人手評価との相関を調べる実験を行う．さらに，自然さが高い英文とそうでない英文のスコアを比較することによって提案手法が適切にスコア付けできているかを検証する実験を行う．

提案手法である 3 つの観点の組合せ手法に加え，個別のコンポーネントである文法性，自然さ，文意の保存の比較を行う．また，従来のリファレンスを用いる評価手法である GLEU および M^2 による評価も行った．リファレンスには，元々の CoNLL2014 のテストデータの正解文 2 文 CoNLL，Sakaguchi らの一般人による正解文 4 文，専門家によるもの 4 文，Bryant と Ng らによる正解文 8 文の計 18 文を用いた．各手法の最終的なスコアは，文単位のスコアの平均によって求めた．

4.1　人手評価との相関による比較実験

提案手法によるスコアと人手評価との相関係数を求める実験を行う．まず，CoNLL 2014 のテストデータと参加チームの出力を用いて従来手法および提案手法で各チームのスコアを求めた．各手法を，Grundkiewicz らによる各文の人手評価から求めた各チームのスコア ([7], Table 3c)　とのスピアマンの順位相関係数 ρ およびピアソンの相関係数 r で評価して比較した．組合せ手法は，式 (1) における文意の保存の重みを $\beta = 0.1$ に固定し，JHLEG データセット [11] における相関が最大となる重みを用いた．

表 1 に各手法の人手との相関を示す．スピアマンの順位相関係数で比較すると，リファレンスレス評価は文法性のみの手法ではリファレンス有りの手法に及ばない．一方，自然さの観点のみで M^2 に匹敵する相関を示した．組合せ手法はリファレンス有りの手法を上回った．

表 1: 人手評価と自動評価の相関係数

手法	ρ	r
リファレンス有り（M^2）	0.648	0.632
リファレンス有り（GLEU）	0.857	0.841
文法性のみ（LT）	0.769	0.641
文法性のみ（Heilman）	0.835	0.759
文意の保存のみ	-0.192	0.198
自然さのみ	0.819	0.864
組合せ（LT）	0.863	0.874
組合せ（Heilman）	0.874	0.878

4.2 自然さが異なる文に対するスコアの比較実験

4.1 節の実験で用いたデータとは性質が異なる文に対しても提案手法が適切に評価できていることを示すための実験を行う．Sakaguchi らが示したように，最小の修正よりも流暢な修正の方が人間の評価が高い．そこで提案手法でも最小の修正より流暢な修正の方に高いスコアをつけることができるかを調べる．

組合せ手法の重みは実験 1 の組合せ手法と同様の重みを用いた．GLEU, M^2 の正解文は CoNLL と Sakaguchi らの一般人によるリファレンスを用いた．

表 2 に人間の訂正に対するスコアを示す．どの手法も原文よりは人間の訂正を高く評価できている．しかしリファレンスを用いる手法は一般人が付与した流暢な修正を正解文に用いているにも関わらず，専門家の流暢さを評価できていない．文法性のみ（LT, Heilman）による評価は最小の編集と流暢な修正に対して同程度のスコアをつけたが，これは最小の編集も文法的である事実を反映している．自然さのみによる評価は明確に流暢な修正の方を高く評価した．

表 2: 人間の訂正に対するスコア．min は専門家 (A，B) による最小の修正，flu は専門家 (A，B) による流暢な修正を表す．アスタリスクは min と flu を比べて統計的有意差があることを示す（$\rho < 0.01$）

	原文	minA	fluA	minB	fluB
M^2	0.00	0.555	0.539*	0.548	0.581*
GLEU	0.517	0.604	0.536*	0.618	0.589*
LT	0.986	0.995	0.996	0.995	0.995
Heilman	0.420	0.485	0.509*	0.488	0.492*
自然さ	0.820	0.898	0.949*	0.908	0.922*
組合せ (LT)	0.864	0.906	0.919*	0.909	0.912*
組合せ (H)	0.788	0.844	0.865*	0.844	0.848*

表 3: 文法性の誤り例

文	Heilman	LT
(i) Do one who suffer from this disease keep it a secret to inform their relatives ?	0.542	1.0
(ii) Does one who suffered from this disease keep it a secret from their relatives ?	0.492	1.0

5　考察

2 つの実験を通して提案手法は従来手法よりも優れた評価性能を示したが，課題もある．課題としては各観点による文単位の評価性能が挙げられる．エラー分析の結果，Heilman モデルのスコアは文レベルで見ると不適切になっている場合があった．表 3 の (i) は非文であり，(ii) は文法的な文であるが，文 (i) の方が高くスコアリングされた．このような問題を解決するためには依存構造の考慮が必要である．例 (i) では主語と動詞の関係を抽出し，数が一致しているかを見る必要がある．また，言語モデルスコアに基づく素性を用いず，他の素性を拡充する方法も考えられる．また，LanguageTool は誤り検出率が低いため多くの文の文法性のスコアが満点になっている．より誤り検出能力の高い検出器を用いれば精度が向上すると考えられる．

また，文意の保存（METEOR）のみによる手法が人手評価と逆相関を示した

のは，内容語が活用語・類義語・スペルミスのいずれにもマッチしない訂正を減点するからであると考えられる．METEOR では類義語の判定を辞書によって行っているが，単語の分散表現による意味類似度を利用すれば改善する可能性がある．誤りを考慮して文意の保存を評価するのは今後の課題である．

提案手法における自然さの評価はエラー分析が困難であったため，学習者の作文の質がわかるデータを用いて自然さの評価精度を検証した．作文の質がわかるデータとして TOEFL11 を用いた．TOEFL11 には作文に対して high, low, middle の 3 段階評価が付与されており，high を付けられた作文は自然さが高いと考えられる．そこで実際に提案手法の自然さのみの観点で学習者の作文を評価した結果を図 1 に示す．作文の人手評価を自然さのスコアだけで予測することはできないが，ある程度関係していることがわかった．自然さのみで作文のレベルを推定することができないのは，作文の人手評価は 1 文の自然さだけではなく，作文の構成や内容の良さも同時に評価しているからであると考えられる．

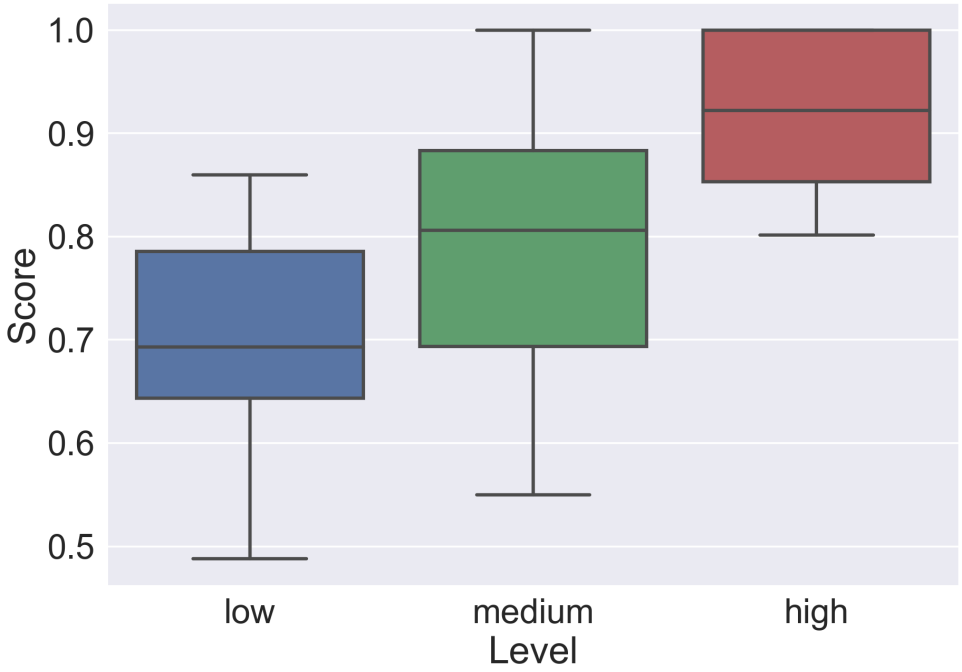

図 1: 自然さに基づくエッセイ評価結果

6 おわりに

本稿では，GEC のための新たなリファレンスレス手法を提案した．従来の GEC システムの評価ではリファレンスを用いた評価尺度が用いられてきた．しかしリファレンスありの評価では人手との相関が高い評価を実現するためには多くのリファレンスが必要であり，リファレンスの作成コストが高い．この問題を解決するために，リファレンスレスの評価を提案した．従来の手法では文法性と文意の保存は考慮されてきたが自然さは考慮されていなかった．そこで，本稿ではその 3 つの観点を組合せた手法を提案し，リファレンスを用いないにも関わらず従来手法よりも人手との相関が高い評価を実現した．また，母国語話者による流暢な英文に対しても適切にスコアリングできていることを示した．今後は，文意の保存の評価方法を変更するなどして評価性能を向上していく予定である．

謝辞

本研究を進めるにあたり，ご指導，ご助言を頂いた乾健太郎教授，岡直観准教授に深く感謝いたします．また，日頃より研究活動を指導してくださいました，水本智也研究特任助教に心より感謝いたします．最後に，日々の議論の中で様々なご助言を頂いた乾・岡研究室の皆様に感謝いたします．

参考文献

[1] Mariano Felice and Ted Briscoe. Towards a standard evaluation method for grammatical error detection and correction. *Proceedings of NAACL-HLT*, pp. 578–587, 2015.

[2] Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. There's No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction. *Proceedings of EMNLP*, pp. 2109–2115, 2016.

[3] Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. Reassessing the Goals of Grammatical Error Correction: Fluency Instead of Grammaticality. *TACL*, pp. 169–182, 2016.

[4] 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 CoNLL Shared Task*, pp. 1–14, 2014.

[5] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. Ground Truth for Grammatical Error Correction Metrics. In *Proceedings of ACL*, pp. 588–593, 2015.

[6] Christopher Bryant and Hwee Tou Ng. How Far are We from Fully Automatic High Quality Grammatical Error Correction? In *Proceedings of ACL*, pp. 697–707, 2015.

[7] Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. Human Evaluation of Grammatical Error Correction Systems. In *Proceedings of EMNLP15*, pp. 461–470, 2015.

[8] Michael Heilman, Joel Tetreault, Aoife Cahill, Nitin Madnani, Melissa Lopez, and Matthew Mulholland. Predicting Grammaticality on an Ordinal Scale. *Proceedings of ACL*, pp. 174–180, 2014.

[9] Jey Han Lau, Alexander Clark, and Shalom Lappin. Unsupervised Prediction of Acceptability Judgements. *Proceedings of ACL*, pp. 1618–1628, 2015.

[10] Michael Denkowski and Alon Lavie. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. *Proceedings of WMT*, pp. 376–380, 2014.

[11] Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. Jfleg: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of EACL*, Valencia, Spain, April 2017. Association for Computational Linguistics.