文法誤り訂正における反復訂正の効果検証

浅野 広樹 †,*1　鈴木 潤 †,*2　水本 智也 *3　乾 健太郎 †,*4
† 東北大学　* 理化学研究所 AIP センター
{[1]asano,[2]jun.suzuki,[4]inui}@ecei.tohoku.ac.jp　[3]tomoya.mizumoto@riken.jp

1 はじめに

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 [2, 6, 9, 10]．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 1）．

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．

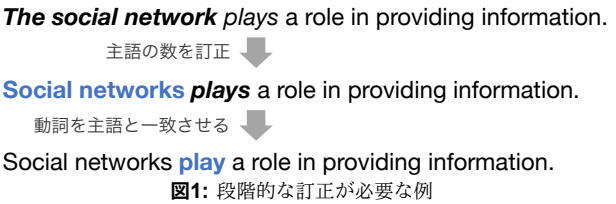

図1: 段階的な訂正が必要な例

2 関連研究

文法誤り訂正の分野は CoNLL-2014 shared task [16] 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 [2] は Convolutional seq2seq [8] 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 [10] は Transformer [17] 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 [6] 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 [11] 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．

表1: 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 反復訂正

Ge ら [6] 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 [6] 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．

4 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 [2] が用いた設定に従って，学習および評価をおこなった．

4.1 データセット

学習には公開されている Lang-8 Learner Corpora [12] と NUS Corpus of Learner English (NUCLE), [5] を用いた．Lang-8 に対しては Chollampatt ら [2] 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

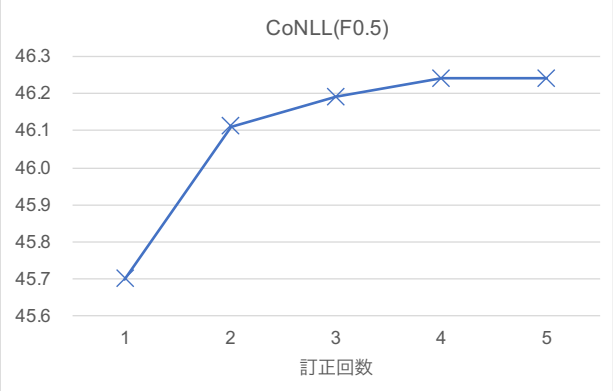

図2: CoNLL 上での反復訂正の性能

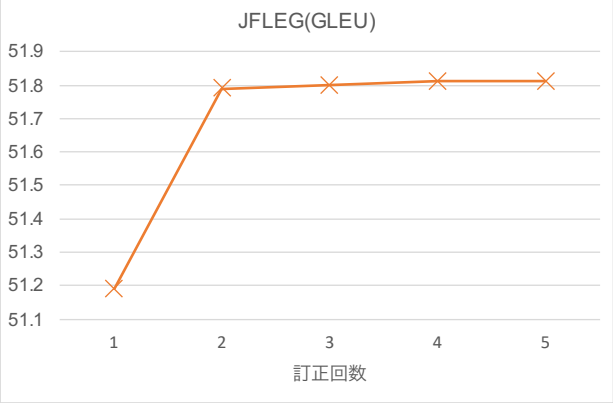

図3: JFLEG 上での反復訂正の性能

を用いた．

4.2 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 [4] で計算された $F_{0.5}$ 値によって評価した[*1]．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) [15] テストセット上において GLEU [13, 14] により評価した．

4.3 モデル設定

訂正モデルには CNN seq2seq [8] を用い，実装は Chollampatt ら [2] 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 [3, 7]．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 [2] が Wikipedia コーパス上で *fastText* [1] 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,

[*1] 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

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．

4.4 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 1に示す．また，図 2, 3に，各反復訂正の評価値を折れ線グラフにより示す．

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 [6] 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 2の obviously 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．

4.5 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 2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 3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．

5 おわりに

本研究は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．今後は本研究で得られた知見を利用して性能

謝辞

本研究の一部は JST CREST(課題番号: JPMJCR1513) の支援を受けて行った．

表2: 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 *obvioualy* 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 .

表3: 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 .

参考文献

[1] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". In: *TACL* (2017), pp. 135–146.

[2] Shamil Chollampatt et al. "A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction". In: *Proceedings of AAAI*. 2018, pp. 5755–5762.

[3] Shamil Chollampatt et al. "Neural Quality Estimation of Grammatical Error Correction". In: *Proceedings of EMNLP*. 2018, pp. 2528–2539.

[4] Daniel Dahlmeier et al. "Better evaluation for grammatical error correction". In: *Proceedings of NAACL*. 2012, pp. 568–572.

[5] Daniel Dahlmeier et al. "Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English". In: *Proceedings of BEA*. 2013, pp. 22–31.

[6] Tao Ge et al. "Fluency Boost Learning and Inference for Neural Grammatical Error Correction". In: *Proceedings of ACL*. 2018, pp. 1055–1065.

[7] Tao Ge et al. "Reaching Human-level Performance in Automatic Grammatical Error Correction: An Empirical Study". In: *CoRR* abs/1807.01270 (2018). arXiv: 1807.01270.

[8] Jonas Gehring et al. "Convolutional Sequence to Sequence Learning". In: *CoRR* abs/1705.03122 (2017). arXiv: 1705.03122.

[9] Roman Grundkiewicz et al. "Near Human-Level Performance in Grammatical Error Correction with Hybrid Machine Translation". In: *Proceedings of NAACL*. 2018, pp. 284–290.

[10] Marcin Junczys-Dowmunt et al. "Approaching Neural Grammatical Error Correction as a Low-Resource Machine Translation Task". In: *Proceedings of NAACL*. 2018, pp. 595–606.

[11] Jared Lichtarge et al. "Weakly Supervised Grammatical Error Correction using Iterative Decoding". In: *CoRR* abs/1811.01710 (2018). arXiv: 1811.01710.

[12] Tomoya Mizumoto et al. "Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners". In: *Proceedings of IJCNLP*. 2011, pp. 147–155.

[13] Courtney Napoles et al. "GLEU Without Tuning". In: *CoRR* abs/1605.02592 (2016). arXiv: 1605.02592.

[14] Courtney Napoles et al. "Ground Truth for Grammatical Error Correction Metrics". In: *Proceedings of ACL-IJCNLP*. 2015, pp. 588–593.

[15] Courtney Napoles et al. "JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction". In: *Proceedings of EACL*. 2017, pp. 229–234.

[16] Hwee Tou Ng et al. "The CoNLL-2014 Shared Task on Grammatical Error Correction". In: *Proceedings of CoNLL*. 2014, pp. 1–14.

[17] Ashish Vaswani et al. "Attention is All you Need". In: *NIPS*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 5998–6008.