文法誤り訂正のコーパス横断評価: 単一コーパス評価で十分か?

三田 雅人 [1,2]　水本 智也 [1]　金子 正弘 [3,1]　永田 亮 [4,1]　乾 健太郎 [2,1]
[1] 理化学研究所 AIP センター　[2] 東北大学　[3] 首都大学東京　[4] 甲南大学
{masato.mita, tomoya.mizumoto}@riken.jp, kaneko-masahiro@ed.tmu.ac.jp,
nagata-nlp2019@hyogo-u.ac.jp., inui@ecei.tohoku.ac.jp

1 はじめに

文法誤り訂正 (Grammatical Error Correction, GEC) は，非母語話者が書いた作文中に含まれる様々な種類の誤りを自動的に訂正するタスクであり，主に第二言語学習のための作文支援を目的に使用されている．文法誤り訂正の主要なアプローチの一つは機械翻訳モデルに基づくものであり，これは，文法誤り訂正タスクを誤った文から正しい文への翻訳タスクと見なして解くアプローチである．実際，代表的なベンチマークデータセットである CoNLL-2014 Shared Task データセット [13] において，現在の state-of-the-art は機械翻訳モデルに基づいた手法により達成されている [2].

一方で，文法誤り訂正の目的として多様な文が入力として入ったとしても頑健に訂正できることを考えた場合，文法誤り訂正の性能評価に関しては不十分な可能性が高い．なぜなら，学習者の習熟度，母語や作文のトピックなどの条件によって様々なバリエーションが入力として想定され，タスクの難易度は各条件下によって異なるにも関わらず，既存研究の多くは単一コーパスを用いた評価を行う傾向があるためである．実際，Grammarly[*1]や Ginger[*2]など，特定のドメインではなく任意のドメインを入力文として想定する文法誤り訂正システムは多い．たとえ提案手法がある一つのコーパス上でベースラインの性能よりも上回ったとしても，別のコーパスで下回るような状況が起きた場合，それは異なる結論をもたらす可能性がある．

そこで本研究では，文法誤り訂正の評価として単一コーパスによる評価は不十分であるという仮説に基づき，コーパス横断評価の必要性について議論する．上記仮説を検証するにあたり，5 つの学習者コーパス (CoNLL-2014, CoNLL-2013 [12], FCE [16], JFLEG [11], KJ [10]) と，3 つのニューラル機械翻訳に基づくモデル (LSTM, CNN, Transformer) および統計的機械翻訳に基づくモデル (SMT) の計 4 つの現在主流となっているモデルを用いて評価実験を行う．評価実験の結果，モデルの順位は各コーパスによって大きく変動することを確認できたため，文法誤り訂正タスクにおいては単一コーパスを用いた評価は不十分であることがわかった．そのため，より頑健な評価を行うための新たな評価手法としてコーパス横断評価を提唱する．

[*1] https://www.grammarly.com/
[*2] http://www.getginger.jp/

2 関連研究

1 節でも述べた通り，既存手法の多くは CoNLL-2014 ベンチマークデータセットを用いて評価されているが，KJ コーパスや JFLEG コーパスといった他のデータセットを用いて評価を行なっている研究も存在する [1, 9]．しかしながら，既存研究はたかだか 2 つのコーパスを用いた評価になっており，本研究のように多様なコーパスを用いたコーパス横断評価を行なっている研究は存在しない．

他分野をみると，例えば構文解析分野では以前は Penn Treebank [7] 上における解析精度を向上させることに重点を置いていた．その後，Penn Treebank に加えて，Ontonotes [4] や Google Web Treebank [14] を含む複数のコーパスを用いて評価を行うようになり，評価手法の頑健性が大幅に改善されている．我々は文法誤り訂正分野においても構文解析分野と同様の状況が起きていると考える．すなわち，現在の文法誤り訂正の評価は CoNLL-2014 ベンチマークデータセットに依存しており，分野全体としてこのデータセット上で過度な開発が行われている恐れがある．

3 実験設定

3.1 評価用コーパス

本研究の目的は，文法誤り訂正の評価として単一コーパスによる評価は不十分であるという仮説に

表1: 各評価コーパスの概要.

コーパス	文数	リファレンス数	単語修正率	習熟度の多様性	トピックの多様性	母語の多様性
CoNLL-2014	1,312	2	12.35%	No	No	No
CoNLL-2013	1,381	1	14.85%	No	No	No
FCE	32,199	1	12.00%	Yes	Yes	Yes
JFLEG	747	4	20.86%	Yes	Yes	Yes
KJ	3,081	1	16.01%	Yes	Yes	No

基づき，コーパス横断評価の必要性について議論することである．そのため，既存研究の多くで使用されている CoNLL-2014 に加えて，CoNLL-2013，FCE，JFLEG，KJ (詳細は下記) の合計 5 つのコーパスを用いて評価実験を行う．コーパス選定にあたり，次の点を考慮した．

- 文法誤り訂正分野で使用されたことのあるコーパスであること
- 学習者の習熟度は誤り分布に大きな影響を及ぼすという仮説に基づき，学習者の習熟度が比較的高い CoNLL-2014 とは反対に，習熟度が比較的低いコーパス (KJ) を加えること

評価実験に用いた各評価用コーパスの詳細を説明する．

CoNLL-2014 [13] : CoNLL-2014 Shared Task の公式データセットであり，既存研究で最もよく使用されている評価用データである．シンガポール国立大学の学生が書いたエッセイを基に作られており，学習者の習熟度は比較的高い．また，エッセイのトピック数は 2 つのみである．

CoNLL-2013 [12] : CoNLL-2013 Shared Task の公式データセットである．CoNLL-2014 と同様に学習者の習熟度は比較的高く，エッセイトピック数は 2 つのみである．

Cambridge ESOL First Certificate in English (FCE) [16] : ケンブリッジ英検の試験答案を基に作られた学習者コーパスである．誤り箇所のタグが付与されているため，文法誤り検出タスクで使用されることが多い．また，様々な国籍の英語学習者が書いたエッセイであるため，学習者の習熟度や母語は多様性がある．なお，エッセイのトピック数は 10 である．

JHU FLUency-Extended GUG corpus (JFLEG) [11] : 様々な国籍の非母語話者が書いた英語答案を基に作られた学習者コーパスである．学習者の習熟度，母語，エッセイトピックともに多様性を担保するよう設計されている [11]．また，母語話者が書くような流暢性のある添削アノテーションが付与されていることも本コーパスの特徴の一つである．そのため，最近の研究では，モデルがどれだけ流暢性のある訂正ができているかを評価するためのベンチマークデータセットとして使用されることが多い．

Konan-JIEM learner corpus (KJ) [10] : 甲南大学の学生が書いたエッセイを基に作られた学習者コーパスである．学習者の習熟度は比較的低い．また，エッセイのトピック数は 10 である．なお，他のコーパスと異なり，綴り誤りの情報と文法誤りの情報が別でアノテーションされていることも特徴の一つである．

各評価用コーパスの特徴を整理したものを表 1に示す．なお，単語修正率は以下の式により算出した．

$$\text{単語修正率} = \frac{\text{単語編集距離の総和}}{\text{コーパスに出現する総単語数}}$$

各コーパスを概観すると，(1) 現在主流である CoNLL-2014 は他のコーパスと比較して学習者の習熟度や母語、トピックに偏りがあること，(2) 一方，本評価実験で選定したコーパスで一定の多様性を確保できることがわかる．

なお，綴り誤りは厳密には文法誤りには含まないが，

- 既存データセットの多くはそれらを区別せずにまとめてアノテーションされている
- また，そのようなデータを用いて訓練した機械翻訳モデルに基づく手法も，それらを区別して訂正できない

という理由から，本評価実験においても従来の慣習に従い綴り誤りを含めて評価を行なう．そのため，綴り誤りと文法誤りが別でアノテーションされている KJ に関しては，統合処理したものを評価実験に使用する．

3.2 訂正モデル

文法誤り訂正モデルの選定については，

- 現在主流であり，かつ複数の誤りを訂正可能な機械翻訳モデルに基づくモデルであること
- CoNLL-2014 上において各モデルは互いに匹敵し合う性能であること

などの要件を踏まえ，3 つの NMT に基づくモデル (LSTM, CNN, Transformer) と SMT に基づくモデルの計 4 つのモデルを実装した．以下に，評価実験に用いた各モデルの詳細について説明する．

LSTM：エンコーダは双方向 LSTM で構成され，デコーダと注意機構は Luong らの seq2seq モデル [6] に準拠している．なお，ハイパーパラメータは Chollampatt ら [1] と同じものを使用した．

CNN：Chollampatt ら [1] の実装[*3]を用いた．なお，ハイパーパラメータは Chollampatt ら [1] と同じものを使用した．

Transformer：Transformer は Vaswani ら [15] によって提案された自己注意型モデルである．本研究では，seq2seq モデルのツール `fairseq` の実装[*4]を用いた．また，ハイパーパラメータは Vaswani ら [15] と同じものを使用した．

SMT：Junczys-Dowmunt ら [5] の実装[*5]を用いた．なお，本評価実験では NMT に基づく各モデルと使用するデータの設定を可能な限り等しくするため，言語モデルの訓練に English Wikipedia を，翻訳モデルの訓練に NUS Corpus of Learner English (NUCLE) および Lang-8 Learner Corpora (Lang-8) を使用した．

3.3 実験設定

各モデル共通の実験設定は次の通りである．訓練データとして，Lang-8 [8] および NUCLE [3] の 2 つのパブリックなデータセットを統合したものを用いた．訓練データセットの前処理および分割は Chollampatt ら [1] に従い，最終的に開発データセットとして NUCLE のサブセット (5.4K)，訓練データとして NUCLE の残りおよび Lang-8 (1.3M) を使用した．評価データセットとしては，3.1 節で述べた計 5 つのコーパスを用いて評価を行った．評価尺度には，$F0.5$ と GLEU を用いた．実験結果として報告するのは 4 つのランダムシードを用いて訓練したモデル性能の平均値である．

[*3] https://github.com/nusnlp/mlconvgec2018
[*4] https://github.com/pytorch/fairseq
[*5] https://github.com/grammatical/baselines-emnlp2016

4 コーパス横断評価

図 1 に $F_{0.5}$ スコアに基づき順位付けされた各モデルの性能を示す．この図から，モデルの性能順位が各コーパスによって大きく変動していることが確認できる．これまで多くの研究が評価に用いていた CoNLL-2014 では，Transformer が 48.62 ポイントと最も高い性能を出している．一方で，FCE や KJ では順位が全体的に大きく変動しており，Transformer は全体で 3 位に転じている．順位が逆転した原因については，学習者の習熟度や母語，エッセイトピックなど様々な要因が複合的に関係していると予想されるため一概に言えない．しかし本評価実験の結果から，CoNLL-2014 上の性能値に基づく議論は特定の条件下でしか成立しない可能性があり，CoNLL-2014 での結果を不用意に一般化してしまうと誤った結論を導く危険性があることがわかった．さらに，図 1 から，モデルの性能はコーパスによって大きく異なることが確認できる．例えば，Transformer の性能は，JFLEG では 60.06 ポイントであるのに対し，CoNLL-2013 では 36.20 ポイントと大きく下回っている．このように，コーパスによって性能は大きく変動するため，見た目の性能値の解釈には十分に注意する必要がある．

次に，GLEU スコアに基づき順位付けされた各モデルの性能を図 2 に示す．GLEU スコアに基づく評価でも $F_{0.5}$ のときと同様の傾向が確認できる．ただし，FCE および KJ におけるモデル順位に関しては，図 1 と図 2 とでは異なる．これは，$F_{0.5}$ と GLEU とではモデルを評価する観点が異なることに起因するものと考えられる．言い換えると，どういうモデルを得たいか，あるいは，モデルの何を評価したいかといった目的意識によって適切に評価データ・評価尺度を設定することが重要であると言える．

1 節において，文法誤り訂正タスクの入力には様々なバリエーションが想定されるため，タスクの難易度は条件によって異なることを述べた．ここでは，想定される入力のバリエーションの一つとして，修正が必要な単語の割合 (単語修正率) の観点で分析する．表 2 は，単語修正率が最も低い場合 (FCE の 62%) と最も高い場合 (JFLEG の 86%) におけるそれぞれのモデル性能を示している．この表から，本評価実験において Precision の高い LSTM は修正率が低いときに，Recall が高い Transformer は修正率が高いときにそれぞれ他のモデルに比べて高い性能を示す傾向を確認できる．このような知見は

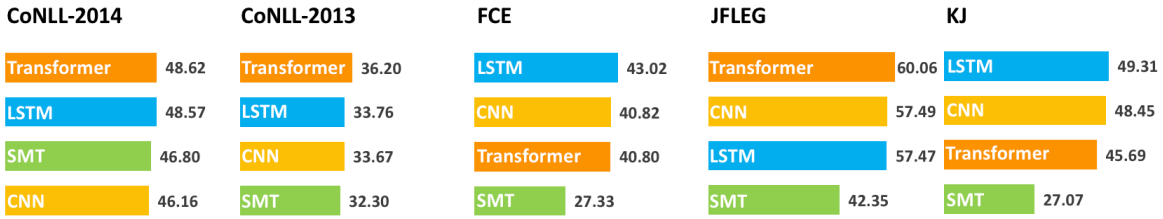

図1: $F_{0.5}$ に基づき順位付けされた各モデルの性能.

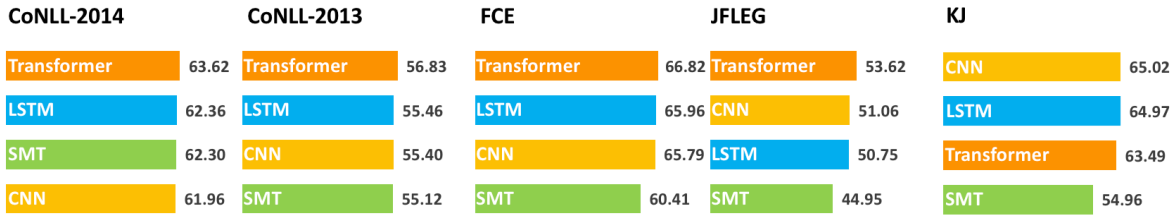

図2: GLEU に基づき順位付けされた各モデルの性能.

表2: 単語修正率観点におけるモデル性能.

単語修正率 (%)	Low (12.00%)			High (20.86%)		
	Precision	Recall	$F_{0.5}$	Precision	Recall	$F_{0.5}$
Transformer	44.93	**29.79**	40.80	67.27	**42.05**	**60.06**
LSTM	**55.48**	21.11	**43.02**	**72.97**	31.09	57.47
CNN	51.27	22.30	40.82	70.85	32.77	57.49
SMT	43.07	10.67	27.33	67.95	16.89	42.35

複数のコーパスで比較評価するコーパス横断評価をしなければ得られない.

5 まとめ

本研究では，文法誤り訂正の評価の方法論について議論した．従来の単一コーパスによる評価がモデルの良し悪しを判断するうえで十分かどうかについて検証するため，5 つの多様なコーパスと 4 つの訂正モデルを選定し，評価実験を行なった．評価実験の結果，モデルの順位は各コーパスによって大きく変動することを確認できたため，入力に多くのバリエーションが想定される文法誤り訂正タスクにおいては単一コーパスを用いた評価は不十分であることがわかった．そのため，より頑健な評価を行うための新たな評価の方法論としてコーパス横断評価を提唱する．

なお，本評価実験では 3.1 節で述べた理由により綴り誤りも含めた設定で評価を行なったが，綴り誤りを含めない文法誤りのみの訂正性能を評価する場合においても本評価実験と同様に見た目の性能値やモデルの順位が大幅に変動する可能性がある．そのため，今後の研究として，今回使用した 5 つのコーパスにおいて綴り誤りの有無がモデルの評価にどのように影響するかについて調査を行う予定である.

参考文献

[1] Shamil Chollampatt and Hwee Tou Ng. “A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction”. In: *AAAI*. 2018, pp. 5755–5762.

[2] Shamil Chollampatt and Hwee Tou Ng. “Neural Quality Estimation of Grammatical Error Correction”. In: *EMNLP*. 2018, pp. 2528–2539.

[3] Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. “Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English”. In: *BEA*. 2013, pp. 22–31.

[4] Eduard Hovy et al. “OntoNotes: The 90% Solution”. In: *NAACL*. 2006, pp. 57–60.

[5] Marcin Junczys-Dowmunt and Roman Grundkiewicz. “Phrase-based Machine Translation is State-of-the-Art for Automatic Grammatical Error Correction”. In: *EMNLP*. 2016, pp. 1546–1556.

[6] Thang Luong, Hieu Pham, and Christopher D. Manning. “Effective Approaches to Attention-based Neural Machine Translation”. In: *EMNLP*. 2015, pp. 1412–1421.

[7] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. “Building a Large Annotated Corpus of English: The Penn Treebank”. In: *Computational Linguistics* 19.2 (1993), pp. 313–330.

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

[9] Tomoya Mizumoto et al. “The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings”. In: *COLING*. 2012, pp. 863–872.

[10] Ryo Nagata, Edward Whittaker, and Vera Sheinman. “Creating a Manually Error-tagged and Shallow-parsed Corpus”. In: *ACL*. 2011, pp. 1210–1219.

[11] Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. “JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction”. In: *EACL*. 2017, pp. 229–234.

[12] Hwee Tou Ng et al. “The CoNLL-2013 Shared Task on Grammatical Error Correction”. In: *CoNLL 2013 Shared Task*. 2013, pp. 1–12.

[13] Hwee Tou Ng et al. “The CoNLL-2014 Shared Task on Grammatical Error Correction”. In: *CoNLL 2014 Shared Task*. 2014, pp. 1–14.

[14] Slav Petrov and Ryan McDonald. “Overview of the 2012 Shared Task on Parsing the Web”. In: *SANCL*. 2012.

[15] Ashish Vaswani et al. “Attention Is All You Need”. In: *NIPS*. 2017, pp. 5998–6008.

[16] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. “A New Dataset and Method for Automatically Grading ESOL Texts”. In: *ACL*. 2011, pp. 180–189.