Vol.2020-NL-246 No.5
Vol.2020-SLP-134 No.5
2020/12/2

英語学習者の母語を考慮した文法誤り訂正のための擬似データ生成

佐藤 義貴[1,a)]　和田 崇史[2,b)]　渡辺 太郎[1,c)]　松本 裕治[3,d)]

概要：本研究では特定の言語を母語にもつ英語学習者が書く文法誤りに頑健な，文法誤り訂正モデルの実現を目指す．文法誤り訂正の研究分野においては，擬似誤りを生成し活用する研究が活発に行われており，多種多様な擬似誤りを生成することができるが，学習者の母語によって誤りの傾向が異なることが考慮されているとは言えない．そこで，本研究では正用文[a]から誤り文を生成する逆翻訳モデルを特定の言語を母語にもつ学習者によって書かれた英文で fine-tune することで，母語の影響を考慮した擬似誤りを生成する手法を提案し，学習者の母語を考慮した文法誤り訂正モデルの性能向上を目指す．

[a] 本稿では，人手で訂正が行われた文を添削文と表記し，単言語コーパスから獲得した文を正用文と表記する．

1. はじめに

文法誤り訂正（Grammatical Error Correction（GEC））は，文法的に誤りを含む文を入力として受け取り，その誤りを訂正した文を出力するタスクである．GEC は誤りを含む文を原言語，正しい文を目的言語とみなし，機械翻訳タスクとして取り組むのが近年では一般的であり，機械翻訳と同様にニューラル Encoder-Decoder モデル [9, 13]（EncDec）が用いられることが多い．

EncDec の特性の 1 つとして，学習済みのモデルを異なるドメインのデータで fine-tune を行うことで，そのドメインに対する性能を向上させることができる点がある．GEC の分野でもその特性を活かし，学習者の習熟度や母語によって誤りの傾向が異なる性質に着目した研究が行われている．例えば Nadejde ら [6] は学習済みの GEC モデルを学習者の習熟度や母語に応じて fine-tune する手法を提案した．しかし，習熟度や母語の情報が付与されたコーパスは一般的なコーパスと比べてデータ量が非常に少なく，GEC モデルを fine-tune する効果も非常に限定的であると考えられる．実際に，Nadejde らの研究では母語ごとに 10,000 文程度の学習者コーパスしか用いることができておらず，加えてそのデータは一般公開されていない．そのため誰しもが実際に利用できるデータ量は各母語について 2,000 文から 3,000 文程度であり，50 万文以上からなる一般的な学習者コーパスと比較すると，非常に少ないと言える．そこで，本研究では正用文から誤り文を生成する「逆翻訳モデル」による擬似エラー生成の手法と，fine-tune による母語適応の手法を組み合わせた手法を提案する．具体的にはまず，添削文から元の誤り文を生成する逆翻訳モデルを学習者コーパスで訓練し，その逆翻訳モデルを母語情報が付与された学習者コーパスで fine-tune する．これにより，母語の影響を考慮した擬似エラーを大量に生成することを目指す．そして最後に，生成された擬似誤りデータと一般的な学習者コーパスを用いて GEC モデルを訓練する．

本研究では，フランス語，日本語，中国語のそれぞれを母語とする英語学習者を対象に実験を行った．その結果，逆翻訳モデルを学習者の母語に適応させることで，全ての母語で GEC モデルの精度改善が見られた．また，提案手法で訓練したモデルを従来法 [6] と同様に母語情報が付与された学習者コーパスで fine-tune することで，精度をさらに改善させることができた．この結果から，GEC モデルを fine-tune する既存手法と，本研究の逆翻訳モデルを fine-tune する手法は補完的であると考えられる．ただし，各母語に適応させたモデル間の精度を比較した結果，擬似エラーが母語ではなくテストデータの特徴に適応しただけの可能性が残されており，今後さらなる検証が必要である．なお，本研究は GEC において 1,000 文にも満たない少量のデータを用いてモデルを学習者の母語に適応することを目指した，最初の研究である．

[1] Nara Institute of Science and Technology
[2] School of Computing and Information Systems, The University of Melbourne
[3] RIKEN Center for Advanced Intelligence Project
[a)] sato.yoshitaka.ss8@is.naist.jp
[b)] twada@student.unimelb.edu.au
[c)] taro@is.naist.jp
[d)] yuji.matsumoto@riken.jp

2. 関連研究

2.1 擬似データ生成に関する研究

逆翻訳は最初 Sennrich ら [7] によってニューラル機械翻訳のための訓練データを拡張する目的で提案され，後に Xie ら [15] がノイズ付き逆翻訳を提案し，GEC のための訓練データ拡張を行った．この研究では，学習者コーパスの添削文から誤りを含む文を生成するようにモデルを訓練し，そのモデルに対し正用文を入力することで擬似的な誤りを含む文を生成する．そしてその生成された擬似誤り文を GEC モデルの訓練データのソース側，生成元となった正用文をターゲット側に含めることでデータ拡張を行う．Zhao ら [16] はモデルの訓練は行わず，正用文に対して単語の置換，削除，追加，及びシャッフルの操作を確率的に行うことで擬似的な誤り文を生成した．Takahashi ら [12] は開発セットにおける英語及びロシア語の学習者の誤り傾向を分析し，その傾向を考慮して擬似データを生成する研究を行った．しかし，これら手法によって生成される擬似的な誤り文は多種多様な誤りを含んでおり，学習者の母語によって誤りの傾向が異なることが考慮されているとは言えない．

2.2 学習者の母語を考慮した研究

Nadejde ら [6] は学習済みの GEC モデルを学習者の母語と習熟度の両方に適応させる手法を提案した．この手法では学習者の習熟度や母語を考慮しない 200 万文の学習者コーパスによって GEC モデルを訓練した後，学習者の母語や習熟度に応じて 10,000 文（fine-tune 用に 8,000 文，開発セットに 2,000 文）の学習者コーパスでモデルを fine-tune することで，その習熟度や母語をもつ学習者によって書かれたテストデータに対する GEC モデルの性能を向上させた．しかし，母語の影響を考慮した GEC の研究はあまり盛んに行われておらず，学習者の母語に焦点を当てた研究にはまだまだ検討の余地があると Wang ら [14] は指摘している．

3. 提案手法

本研究では学習済みの逆翻訳モデルを特定の言語を母語にもつ学習者によって書かれた英文で fine-tune することで，母語の影響を考慮した擬似誤りを生成する手法を提案する．まずはじめに，学習者コーパスの添削文をソース側，誤りを含む文をターゲット側として逆翻訳モデルの事前訓練を行う．その後，特定の言語を母語にもつ学習者によって書かれた学習者コーパスで逆翻訳モデルを fine-tune し，母語の影響をより考慮した擬似誤り文を生成するモデルを訓練する．そして，上記の方法によって得られたモデルに対して正用文を入力し，擬似誤り文を生成する．最後に，生成された擬似誤り文をソース側，正用文をターゲット側のデータに加え，GEC モデルの訓練を行う．なお，Xie ら [15] による手法では逆翻訳モデルの生成を行う際に，幅広い種類の誤りを生成する目的でビームサーチのスコアにペナルティを与える操作を行っていたが，本研究では多様な誤りを生成するのではなく母語を考慮した誤りを生成したいため，デコード時の単語の生成確率の分布に応じてサンプリングする手法 [7] を選択した．[*1]

表 1 データの分割

	データセット	文（ペア）の数
学習者コーパス	BEA-train	543,783
	BEA-valid	4,384
	FCE-valid	663（different）
		337（same）
	FCE-test	1,000
単言語コーパス	wikipedia	1,400,000

4. 実験

4.1 データセット

表 1 に本実験のデータセットの文（ペア）数を示す．実験では BEA2019 で開催された GEC の SharedTask で公開されたデータセットを使用する．Lang8 コーパス [5]，NUCLE コーパス [3]，W&I+LOCNESS コーパス [1,10] を BEA-train と BEA-valid に割り当てる．ただし，BEA-train はソースとターゲットが同一のペアを削除し，BEA-valid は逆翻訳モデルの開発セットのために使用する．そして，学習者の母語の情報が付与されている FCE コーパスから学習者の母語ごとに 2,000 文ずつ抽出し FCE-valid と FCE-test に 1,000 文ずつ割り当てる．分割するプロセスは冒頭の 1,000 文を FCE-test として，残りのデータから誤り文と添削文が非同一な文を 663 文，同一な文 337 文を組み合わせたものを FCE-valid ととする．FCE-vaild 内の非同一な 663 文を fine-tune の際にも使用する．擬似誤り文の生成元コーパスには wikipedia 英語コーパス[*2]から抽出した 140 万文を使用する．

4.2 実験設定

実験では

- BEA-train で訓練した GEC モデル
- BEA-train と通常の逆翻訳モデルで生成した 140 万文の擬似データで訓練した GEC モデル

これらのモデルとの比較を行う．性能の評価には ERRANT [2] を用いる．fine-tune と評価にはフランス語，日

[*1] Kiyono ら [4] の研究では 140 万文の擬似データを使った場合はサンプリングの手法とペナルティを与える手法との差はほとんど生じていない．
[*2] WMT20 で公開されているデータを用いた．http://data.statmt.org/wmt20/translation-task/ps-km/wikipedia.en.lid_filtered.test_filtered.xz

表 2　学習者コーパスに擬似エラーを加えて訓練したモデルの性能の比較．母語の影響を考慮するために，提案手法では通常の逆翻訳モデルを母語情報付きの学習者コーパスで fine-tune している．

		学習者の母語								
		フランス語			日本語			中国語		
訓練データ	擬似データ	*Prec.*	*Rec.*	$F_{0.5}$	*Prec.*	*Rec.*	$F_{0.5}$	*Prec.*	*Rec.*	$F_{0.5}$
BEA-train	なし	43.2	**24.3**	37.4	42.8	32.6	40.3	44.5	27.6	39.6
	通常の逆翻訳モデル	49.2	22.8	39.9	44.9	32.0	41.5	48.8	25.4	41.2
	提案手法	**49.5**	23.8	**40.7**	**45.7**	**34.2**	**42.8**	**52.5**	**29.0**	**45.2**

表 3　GEC モデルをそれぞれ学習した後，そのモデルをさらに母語情報付きの学習者コーパスで fine-tune したときの結果（括弧内の数字は上昇幅を示す）．

		学習者の母語								
		フランス語			日本語			中国語		
訓練データ	擬似データ	*Prec.*	*Rec.*	$F_{0.5}$	*Prec.*	*Rec.*	$F_{0.5}$	*Prec.*	*Rec.*	$F_{0.5}$
BEA-train	なし	**52.0**	23.3	41.7 (+4.3)	**50.6**	29.6	44.3 (+4.0)	54.8	27.3	45.6 (+6.0)
	通常の逆翻訳モデル	49.0	24.3	40.7 (+0.8)	45.8	33.4	42.7 (+1.2)	52.7	29.9	45.7 (+4.5)
	提案手法	51.8	**25.3**	**42.8** (+2.1)	49.9	**35.4**	**46.1** (+3.3)	**55.5**	**31.2**	**48.0** (+2.8)

本語，中国語のそれぞれを母語にもつ学習者によって書かれた英文を使う．

4.3　GEC モデルと逆翻訳モデル

GEC モデル，逆翻訳モデルの両方とも，既存の EncDec モデルとして，Transformer [13] を用いた．具体的には Vaswani ら [13] の定義した”Transformer（big）”設定を用いた．fairseq[*3]の実装を使って実験を行った．入出力はバイト対符号化 [8]（BPE）を用いてサブワード化し，語彙サイズは 8,000 とした．BPE コードは GEC モデルでは添削文側から，逆翻訳モデルでは誤り文側から獲得したものを使用した．

GEC モデルは誤り文から添削文を生成するように訓練を行う．開発セットには母語に応じた FCE-valid を使用する．

逆翻訳モデルは添削文から誤り文を生成するように訓練を行う．開発セットには BEA-valid を使用する．逆翻訳モデルの fine-tune を行う際のハイパーパラメータ（エポック数，ドロップアウト率，ミニバッチ内の最大トークン数）は，スペイン語を母語にもつ学習者によって書かれた FCE-valid と FCE-test を用いて決定した．[*4]

4.4　実験結果

表 2 に各設定で学習者の母語それぞれに対して訓練を行ったモデルの FCE-test に対する性能を ERRANT で評価した際のスコアを示す．通常の逆翻訳モデルによって生成された擬似データを訓練に用いた場合よりも，提案手法によって母語ごとに生成された擬似データを訓練に用いた場合の方がそれぞれの母語についてモデルの性能（$F_{0.5}$）が向上することを確認した．

4.5　GEC モデルの fine-tune

母語に応じたデータで GEC モデルの fine-tune を行う従来法 [6] と同様に，我々の手法で獲得したモデルに対しても FCE-valid を使って fine-tune を行った結果を表 3 に示す．fine-tune を行う際は，スペイン語を母語にもつ学習者によって書かれた FCE-valid で fine-tune を行った際に，スペイン語を母語にもつ学習者によって書かれた FCE-test に対して最も高い $F_{0.5}$ を達成できたハイパーパラメータを使う．その結果，663 文という少ないデータにも関わらず，先行研究と同様に全ての訓練設定，母語で精度が改善された．そして，その中でも提案手法のモデルが学習者の母語それぞれに対して最も高い精度を達成したことから，GEC モデルの fine-tune と提案手法は補完的であると言える．

表 4　異なる母語のデータを評価した場合の比較（$F_{0.5}$）

		学習者の母語		
		フランス語	日本語	中国語
モデル	フランス語モデル	40.7	43.0	42.2
	日本語モデル	40.4	42.8	42.1
	中国語モデル	**41.1**	**44.1**	**45.2**

4.6　異なる母語のデータに対する性能

フランス語，日本語，中国語のそれぞれを母語にもつ学習者のデータに対し提案手法の有効性が確認できたが，これは単に擬似誤り文の特徴が FCE というドメインに対しての適応に成功しただけという可能性がある．その検証を行うため，我々の提案手法で獲得したモデルで異なる母語をもつ学習者によって書かれた FCE-test に対するスコア

[*3] https://github.com/pytorch/fairseq
[*4] FCE-test の label smoothed cross entropy [11] が最も低くなった値を用いた．

を計測した．その結果を表 4 に示す．表より，中国語を母語にもつ学習者に特化したモデルでフランス語及び日本語のデータに対しての精度を計測したところ，それぞれの母語に特化したモデルよりも上回る結果が得られた．

このことから，本実験では提案手法による母語への適応度合いは低く，母語への適応よりも FCE へのドメインへの適応が精度向上に寄与した可能性が高い．しかし，通常の逆翻訳モデルを用いた場合よりもフランス語，日本語，中国語を母語にもつ学習者によって書かれた FCE-test に対してのスコアは上昇しており，また本実験では fine-tune に用いた文数が 663 文と極めて少ないため，データを増やすことで母語適応がより上手くいく可能性もある．

5. おわりに

本研究では，従来法の逆翻訳モデルによる擬似データ生成と，GEC モデルを特定の言語を母語にもつ学習者によって書かれた文で fine-tune を行う手法を発展させ，逆翻訳モデルの fine-tune を行った．実験の結果，通常の逆翻訳モデルを用いた場合よりもテストデータに対する性能が向上することが確認できた．ただ，これは特定の言語を母語にもつ学習者の誤りに頑健になったのではなく，FCE というドメインに対して頑健になっただけの可能性もある．これは，開発セットのデータ数の少なさによるモデルの最適化の失敗及び，テストセットのデータ数の少なさによる性能評価の不確かさも影響していると考えられる．しかし，母語ごとに 663 文という極めて少量のデータにもかかわらず逆翻訳モデルの生成結果を特定のドメイン（FCE）に適応できた結果でもあるので，今後は，fine-tune に使うデータの量を増やすなどして逆翻訳モデルの生成をさらに学習者の誤り傾向に適応させることに取り組んでいく．また逆翻訳モデルによる擬似データ生成は，擬似データを増やすことで性能が上がることが示されているため，我々の提案手法をさらに大規模なデータに対して用い，その効果を検証していく．

参考文献

[1] Bryant, C., Felice, M., Andersen, Ø. E. and Briscoe, T.: The BEA-2019 Shared Task on Grammatical Error Correction, *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Florence, Italy, Association for Computational Linguistics, pp. 52–75 (online), DOI: 10.18653/v1/W19-4406 (2019).

[2] Bryant, C., Felice, M. and Briscoe, T.: Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, Association for Computational Linguistics, pp. 793–805 (online), DOI: 10.18653/v1/P17-1074 (2017).

[3] Dahlmeier, D., Ng, H. T. and Wu, S. M.: Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English, *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, Association for Computational Linguistics, pp. 22–31 (online), available from ⟨https://www.aclweb.org/anthology/W13-1703⟩ (2013).

[4] Kiyono, S., Suzuki, J., Mita, M., Mizumoto, T. and Inui, K.: An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Association for Computational Linguistics, pp. 1236–1242 (online), DOI: 10.18653/v1/D19-1119 (2019).

[5] Mizumoto, T., Komachi, M., Nagata, M. and Matsumoto, Y.: Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners, *Proceedings of 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, Asian Federation of Natural Language Processing, pp. 147–155 (online), available from ⟨https://www.aclweb.org/anthology/I11-1017⟩ (2011).

[6] Nadejde, M. and Tetreault, J.: Personalizing Grammatical Error Correction: Adaptation to Proficiency Level and L1, *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, Hong Kong, China, Association for Computational Linguistics, pp. 27–33 (online), DOI: 10.18653/v1/D19-5504 (2019).

[7] Sennrich, R., Haddow, B. and Birch, A.: Improving Neural Machine Translation Models with Monolingual Data, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 86–96 (online), DOI: 10.18653/v1/P16-1009 (2016).

[8] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 1715–1725 (online), DOI: 10.18653/v1/P16-1162 (2016).

[9] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, *Advances in Neural Information Processing Systems 27* (Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. and Weinberger, K. Q., eds.), Curran Associates, Inc., pp. 3104–3112 (online), available from ⟨http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf⟩ (2014).

[10] Sylviane, G.: The computer learner corpus: A versatile new source of data for SLA research, *Learner English on Computer* (Sylviane, G., ed.), London and New York, Addison Wesley Longman, pp. 3–18 (1998).

[11] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: Rethinking the Inception Architecture for Computer Vision, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,*, (online), available from ⟨http://arxiv.org/abs/1512.00567⟩ (2016).

[12] Takahashi, Y., Katsumata, S. and Komachi, M.: Grammatical Error Correction Using Pseudo Learner Corpus Considering Learner's Error Tendency, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Online, Association for Computational Linguistics, pp. 27–

32 (online), DOI: 10.18653/v1/2020.acl-srw.5 (2020).

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems 30* (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds.), Curran Associates, Inc., pp. 5998–6008 (online), available from ⟨http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf⟩ (2017).

[14] Wang, Y., Wang, Y., Liu, J. and Liu, Z.: A Comprehensive Survey of Grammar Error Correction (2020).

[15] Xie, Z., Genthial, G., Xie, S., Ng, A. and Jurafsky, D.: Noising and Denoising Natural Language: Diverse Back-translation for Grammar Correction, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, Association for Computational Linguistics, pp. 619–628 (online), DOI: 10.18653/v1/N18-1057 (2018).

[16] Zhao, W., Wang, L., Shen, K., Jia, R. and Liu, J.: Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, Association for Computational Linguistics, pp. 156–165 (online), DOI: 10.18653/v1/N19-1014 (2019).

正誤表

下記の箇所に誤りがございました．お詫びして訂正いたします．

訂正箇所	誤	正
2 ページ目 右側 6-7 行目	サンプリングする手法[7]を選択した．	サンプリングする手法を選択した．