事前学習済みマスク言語モデルを組み合わせた 符号化復号化文法誤り訂正モデル

金子 正弘^{†,††}

1 はじめに

本稿では ACL2020 に採択された “Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction” (Kaneko, Mita, Kiyono, Suzuki, and Inui 2020) について解説を行う。普段、著者は東京都立大学小町研究室にて研究を行っているが、この研究は 2019 年 11 月に東北大学乾・鈴木研究室に滞在した際に行った研究である。

2 背景

論文の解説を行う前に、まず文法誤り訂正 (Grammatical Error Correction: GEC) の動向について説明を行う。GEC は文法的に誤っている可能性がある文 (誤文) が入力として与えられ、文法的に正しい文 (正文) に訂正するタスクである。GEC の技術は Grammarly¹ のような言語教育・学習支援サービスとして広く活用されている。国外では ACL2019 の併設ワークショップにおいて 24 チームが参加した共通タスク²が開催され、国内では言語処理年次大会 2020 において 13 件の発表が行われるなど、ここ数年 GEC は国内外で大きな盛り上がりを見せている。

近年の GEC の傾向として、生コーパスから疑似的な文法誤りデータを作成し GEC モデルを学習する手法が主流となっている。一般的に GEC の学習に用いられているデータは 100 万文対ほどであり、数千万文の疑似データを作成し学習することで劇的に訂正性能を改善することが可能となる。GEC における疑似データの研究は大きく分けて 2 つある。(1) 機械翻訳の逆翻訳を基に正文から誤文を生成するモデルを学習し疑似データを生成する手法 (Xie, Genthial, Xie, Ng, and Jurafsky 2018; Kiyono, Suzuki, Mita, Mizumoto, and Inui 2019) や (2) 削除、挿入、置換、入れ替えやマスクなどのランダムな操作により作成する手法 (Zhao, Wang, Shen, Jia, and Liu 2019; Grundkiewicz, Junczys-Dowmunt, and Heafield 2019; Takahashi, Katsumata, and Komachi 2020) などが研究されている。特に、疑似データを用いてモデルを事前学習すること

[†] 東京都立大学

^{††} 理化学研究所

¹ <https://www.grammarly.com/>

² <https://www.cl.cam.ac.uk/research/nl/bea2019st/>

が有効であることがわかっている (Kiyono et al. 2019).

大規模データを活用する研究として他にも, BERT (Devlin, Chang, Lee, and Toutanova 2019) のような事前学習されたマスク言語モデル (Masked Language Model: MLM) を使う手法が提案されている. MLM を用いたリランキングを行う手法 (Kaneko, Hotate, Katsumata, and Komachi 2019; Chollampatt, Wang, and Ng 2019), 事前に MLM により正文をフィルタリングする手法 (Asano, Mita, Mizumoto, and Suzuki 2019) や MLM 自体を GEC モデルとして用いる研究 (Kantor, Katz, Choshen, Cohen-Karlik, Liberman, Toledo, Menczel, and Slonim 2019; Awasthi, Sarawagi, Goyal, Ghosh, and Piratla 2019) などがある.

3 マスク言語モデルを組み込んだ文法誤り訂正

この章から ACL2020 に採択された著者の論文について解説を行う. 本研究では GEC における符号化復号化モデルに事前学習された MLM を組み込む研究を行った. MLM は分類や系列ラベリングタスクなどで大幅な改善をもたらすことが示されている. さらに, 最近では MLM と符号化復号化モデルを組み合わせる研究が行われており, 機械翻訳や要約などの系列変換タスクで成功を収めている. GEC では符号化復号化モデルを用いることが主流であり, 現在の最先端の結果のほとんどは符号化復号化モデルにより達成されている (Lichtarge, Alberti, Kumar, Shazeer, Parmar, and Tong 2019; Kiyono et al. 2019). この傾向に照らして, GEC の符号化復号化モデルは他の系列変換タスクと同様に MLM の恩恵を受けることができるのか? という 1 つの疑問が生じる. GEC では他のタスクとは異なり入力テキストに文法誤りや流暢ではない表現が含まれている. MLM は文法誤りや非流暢な表現を含まないテキストから学習されているため, 既存手法をそのまま適応し GEC でうまくいくかは明らかではない.

MLM を符号化復号化モデルに活用する手法として, 主に初期化手法 (init) と素性手法 (fuse) の 2 つがある. **init**: 符号化復号化モデルの重みを事前学習された MLM の重みで初期化し, 対象タスクのデータでファインチューニングする (Lample and Conneau 2019). この手法の問題点は, GEC のような系列変換タスクでは大規模なデータが学習に必要なため, 初期化された情報に対して破壊的忘却を引き起こし, 事前学習された情報を十分に活用することができない可能性があることである. そして, この手法は 2 章で述べた疑似データを用いた事前学習手法とどちらも初期化する必要があるため併用できない. **fuse**: 誤文を入力とする MLM の出力を追加の素性として誤文と一緒に符号化復号化モデルに与える (Zhu, Xia, Wu, He, Qin, Zhou, Li, and Liu 2020). この手法は init とは違い MLM の重みを更新しないため, 事前学習された情報が保持され破壊的忘却は発生しない. 一方で, MLM は学習者が書いた文法誤りや非流暢なテキストなどの GEC タスク固有の情報を考慮することができないことが問題である.

著者は既存手法の問題点を解決するために, 入力テキストの文法的に誤った箇所を検出する

文法誤り検出 (Grammatical Error Detection: GED) タスクによりファインチューニングされた MLM を用いる素性手法 (fuse GED) を提案した. **fuse GED**: まず事前学習された MLM を GED でファインチューニングする. そして, GED でファインチューニングされた MLM の出力を fuse 手法と同じように素性として用いる. GED は GEC のサブタスクであり文法誤りや流暢ではない表現などを学習することができる. さらに, GED は系列ラベリングタスクであるため, 小規模なデータで十分に学習できるため破壊的忘却を引き起こす可能性が低い. そして, 素性として用いるため疑似データで事前学習された GEC モデルと併用することが可能である.

4 結果

表 1 はベースライン, 既存手法と提案手法の GEC の結果を示している. MLM として BERT (Devlin et al. 2019) を用いて実験を行った. 学習・開発データとして BEA データ (Bryant, Felice, Andersen, and Briscoe 2019) を用いた. 評価データとして BEA-test, CoNLL-14 (Ng, Wu, Briscoe, Hadiwinoto, Susanto, and Bryant 2014) と JFLEG (Napoles, Sakaguchi, and Tetreault 2017) を用い, 評価指標としては ERRANT (Bryant, Felice, and Briscoe 2017), M^2 (Dahlmeier and Ng 2012) と GLEU (Napoles, Sakaguchi, Post, and Tetreault 2015) を用いた. 1 つ目のグループは疑似データを用いない単一モデルの結果, 2 つ目のグループは疑似データを用いた単一モデルの結果, 3 つ目のグループは疑似データを用いたアンサンブルの結果である.

まず, 1 つ目の疑似データを用いないグループでは, MLM を用いるどの手法も MLM を使わない手法と比較して性能が大幅に改善されている. 特に提案手法である BERT-fuse GED は全

表 1 各モデルの GEC の結果

	BEA-test (ERRANT)			CoNLL-14 (M^2)			JFLEG
	P	R	F _{0.5}	P	R	F _{0.5}	GLEU
w/o BERT	51.5	43.2	49.6	59.2	31.2	50.2	52.7
BERT-init	55.1	43.7	52.4	61.3	31.5	51.4	53.0
BERT-fuse	57.5	44.9	54.4	62.3	31.3	52.0	54.1
BERT-fuse GED	58.1	44.8	54.8	63.6	33.0	53.6	54.4
w/o BERT	66.1	59.9	64.8	68.5	44.8	61.9	61.0
BERT-fuse	66.6	60.0	65.2	68.3	45.7	62.1	61.2
BERT-fuse GED	67.1	60.1	65.6	69.2	45.6	62.6	61.3
Lichtarge et al. 2019	—	—	—	65.5	37.1	56.8	61.6
Kiyono et al. 2019	65.5	59.4	64.2	67.9	44.1	61.3	59.7
BERT-fuse GED + R2L	72.3	61.4	69.8	72.6	46.4	65.2	62.0
Lichtarge et al. 2019	—	—	—	66.7	43.9	60.4	63.3
Kiyono et al. 2019	74.7	56.7	70.2	72.4	46.1	65.0	61.4

太字は各グループで最も高い結果を示している. (Kaneko et al. 2020)

てのデータセットで最も良い結果を示している。2つ目のグループの結果から、疑似データを用いた場合でも MLM を用いることでさらに性能が向上することがわかる。そして、単一モデルで BEA-test と CoNLL-14 の2つデータで世界最高精度を達成している。3つ目のグループではアンサンブルを行うことで CoNLL-14 において提案手法が世界最高精度を達成している。これらの結果から、GEC において MLM を用いることは有効であり、特に GEC のサブタスクである GED でファインチューニングすることでより効果的に MLM を活用できることを明らかにした。

5 おわりに

本稿では MLM を GEC の符号化復号化モデルと組み合わせる手法を提案し ACL2020 に採択された論文の解説を行った。

参考文献

- Asano, H., Mita, M., Mizumoto, T., and Suzuki, J. (2019). “The AIP-Tohoku System at the BEA-2019 Shared Task.” In *BEA*, pp. 176–182.
- Awasthi, A., Sarawagi, S., Goyal, R., Ghosh, S., and Piratla, V. (2019). “Parallel Iterative Edit Models for Local Sequence Transduction.” In *EMNLP-IJCNLP*, pp. 4260–4270.
- Bryant, C., Felice, M., Andersen, Ø. E., and Briscoe, T. (2019). “The BEA-2019 Shared Task on Grammatical Error Correction.” In *BEA*, pp. 52–75.
- Bryant, C., Felice, M., and Briscoe, T. (2017). “Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction.” In *ACL*, pp. 793–805.
- Chollampatt, S., Wang, W., and Ng, H. T. (2019). “Cross-Sentence Grammatical Error Correction.” In *ACL*, pp. 435–445.
- Dahlmeier, D. and Ng, H. T. (2012). “Better Evaluation for Grammatical Error Correction.” In *NAACL*, pp. 568–572.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In *NAACL*, pp. 4171–4186.
- Grundkiewicz, R., Junczys-Dowmunt, M., and Heafield, K. (2019). “Neural Grammatical Error Correction Systems with Unsupervised Pre-training on Synthetic Data.” In *BEA*, pp. 252–263.
- Kaneko, M., Hotate, K., Katsumata, S., and Komachi, M. (2019). “TMU Transformer System Using BERT for Re-ranking at BEA 2019 Grammatical Error Correction on Restricted Track.” In *BEA*, pp. 207–212.

- Kaneko, M., Mita, M., Kiyono, S., Suzuki, J., and Inui, K. (2020). “Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction.” In *ACL*, pp. 4248–4254.
- Kantor, Y., Katz, Y., Choshen, L., Cohen-Karlik, E., Liberman, N., Toledo, A., Menczel, A., and Slonim, N. (2019). “Learning to combine Grammatical Error Corrections.” In *BEA*, pp. 139–148.
- Kiyono, S., Suzuki, J., Mita, M., Mizumoto, T., and Inui, K. (2019). “An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction.” In *EMNLP-IJCNLP*, pp. 1236–1242.
- Lample, G. and Conneau, A. (2019). “Cross-lingual Language Model Pretraining.” In *NeurIPS*, pp. 7059–7069.
- Lichtarge, J., Alberti, C., Kumar, S., Shazeer, N., Parmar, N., and Tong, S. (2019). “Corpora Generation for Grammatical Error Correction.” In *NAACL*, pp. 3291–3301.
- Napoles, C., Sakaguchi, K., Post, M., and Tetreault, J. (2015). “Ground Truth for Grammatical Error Correction Metrics.” In *NAACL*, pp. 588–593.
- Napoles, C., Sakaguchi, K., and Tetreault, J. (2017). “JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction.” In *EACL*, pp. 229–234.
- 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 *CoNLL*, pp. 1–14.
- Takahashi, Y., Katsumata, S., and Komachi, M. (2020). “Grammatical Error Correction Using Pseudo Learner Corpus Considering Learner’s Error Tendency.” In *ACL SRW*, pp. 27–32.
- Xie, Z., Genthial, G., Xie, S., Ng, A., and Jurafsky, D. (2018). “Noising and Denoising Natural Language: Diverse Backtranslation for Grammar Correction.” In *NAACL*, pp. 619–628.
- Zhao, W., Wang, L., Shen, K., Jia, R., and Liu, J. (2019). “Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data.” In *NAACL*, pp. 156–165.
- Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., and Liu, T. (2020). “Incorporating BERT into Neural Machine Translation.” In *ICLR*.

略歴

金子 正弘：2016 年北見工業大学工学部情報システム工学科卒業。2018 年首都大学東京システムデザイン研究科情報通信システム学域修士課程修了。同年、東京都立大学システムデザイン研究科情報科学域博士課程に進学、現在に至る。