Stronger Baselines for Grammatical Error Correction Using a Pretrained Encoder–Decoder Model

Satoru Katsumata[*] and **Mamoru Komachi**
Tokyo Metropolitan University
satoru.katsumata@retrieva.jp, komachi@tmu.ac.jp

Abstract

Studies on grammatical error correction (GEC) have reported the effectiveness of pretraining a Seq2Seq model with a large amount of pseudodata. However, this approach requires time-consuming pretraining for GEC because of the size of the pseudodata. In this study, we explore the utility of bidirectional and auto-regressive transformers (BART) as a generic pretrained encoder–decoder model for GEC. With the use of this generic pretrained model for GEC, the time-consuming pretraining can be eliminated. We find that monolingual and multilingual BART models achieve high performance in GEC, with one of the results being comparable to the current strong results in English GEC. Our implementations are publicly available at GitHub[1].

1 Introduction

Grammatical error correction (GEC) is the automatic correction of grammatical and other language-related errors in text. Most works regard this task as a translation task and use encoder–decoder (Enc–Dec) architectures to convert ungrammatical sentences to grammatical ones. This Enc–Dec approach often does not require linguistic knowledge of the target language. Strong Enc–Dec models for GEC are pretrained with a large amount of artificially generated data, commonly referred to as 'pseudodata', that is created by introducing artificial error to a monolingual corpus. Hereafter, pretraining using pseudodata aimed at the GEC task is referred to as **task-oriented** pretraining (Kiyono et al., 2019; Grundkiewicz et al., 2019; Náplava and Straka, 2019; Kaneko et al., 2020). For example, Kiyono et al. (2019) generated a pseudo corpus using back-translation and achieved strong results for English GEC. Náplava and Straka (2019) generated a pseudo corpus by introducing artificial errors into monolingual corpora and achieved the best scores for GEC in several languages by adopting the methods proposed by Grundkiewicz et al. (2019).

These task-oriented pretraining approaches require extensive use of a pseudo-parallel corpus. Specifically, Grundkiewicz et al. (2019) used 100M ungrammatical and grammatical sentence pairs, while Kiyono et al. (2019) and Kaneko et al. (2020) used 70M sentence pairs, which required time-consuming pretraining of GEC models using the pseudo corpus.

In this study, we determined the effectiveness of publicly available pretrained Enc–Dec models for GEC. Specifically, we investigated pretrained models without the need for pseudodata. We explored a pretrained model proposed by Lewis et al. (2020) called bidirectional and auto-regressive transformers (BART). Liu et al. (2020) also proposed multilingual BART. These models were pretrained by predicting the original sequence, given a masked and shuffled sentence. The motivation for using these models for GEC was that it achieved strong results for several text generation tasks, such as summarization; we refer to it as a **generic** pretrained model.

We used generic pretrained BART models to compare with GEC models using a pseudo-corpus approach (Kiyono et al., 2019; Kaneko et al., 2020; Náplava and Straka, 2019). We conducted GEC experiments for four languages: English, German, Czech, and Russian. The Enc–Dec model based on BART achieved results comparable with those of current strong Enc–Dec models for English GEC. The multilingual model also showed high performance in other languages, despite only requiring fine-tuning. These results suggest that BART can be used as a simple baseline

[*] Currently working at Retrieva, Inc.

[1] https://github.com/Katsumata420/generic-pretrained-GEC

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 827–832
December 4 - 7, 2020. ©2020 Association for Computational Linguistics

for GEC.

2 Previous Work

The Enc–Dec approach for GEC often uses the task-oriented pretraining strategy. For example, Zhao et al. (2019) and Grundkiewicz et al. (2019) reported that pretraining of the Enc–Dec model using a pseudo corpus is effective for the GEC task. In particular, they introduced word- and character-level errors into a sentence in monolingual corpora. They developed a confusion set derived from a spellchecker and randomly replaced a word in a sentence. They also randomly deleted a word, inserted a random word, and swapped a word with an adjacent word. They performed these same operations, i.e., replacing, deleting, inserting, and swapping, for characters. The pseudo corpus made by the above methods consisted of 100M training samples. Our study aims to investigate whether the generic pretrained models are effective for GEC, because pretraining with such a large corpus is time-consuming.

Náplava and Straka (2019) adopted Grundkiewicz et al. (2019)'s method for several languages, including German, Czech, and Russian. They trained a Transformer (Vaswani et al., 2017) with pseudo corpora (10M sentence pairs), and achieved current state-of-the-art (SOTA) results for German, Czech, and Russian GEC. We compared their results with those of the generic pretrained model to confirm whether the model was effective for GEC in several languages.

Kiyono et al. (2019) explored the generation of a pseudo corpus by introducing random errors or using back-translation. They reported that a task-oriented pretraining with back-translation data and character errors is better than that with pseudodata based on random errors. Kaneko et al. (2020) combined Kiyono et al. (2019)'s pretraining approach with BERT (Devlin et al., 2019) and improved Kiyono et al. (2019)'s results. Specifically, Kaneko et al. (2020) fine-tuned BERT with a grammatical error detection task. The fine-tuned BERT outputs for each token were combined with the original tokens as a GEC input. Their study is similar to our research in that both studies use publicly available generic pretrained models to perform GEC. The difference between these studies is that Kaneko et al. (2020) used the architecture of the pretrained model as an encoder. Therefore, their method still requires pretraining with a large amount of pseudodata.

The current SOTA approach for English GEC uses the sequence tagging model proposed by Omelianchuk et al. (2020). They designed token-level transformations to map input tokens to target corrections to produce training data. The sequence tagging model then predicts the transformation corresponding to the input token. We do not attempt to make a comparison with this approach, as the purpose of our study is to create a strong GEC model without using pseudodata or linguistic knowledge.

3 Generic Pretrained Model

BART (Lewis et al., 2020) is pretrained by predicting an original sequence, given a masked and shuffled sequence using a Transformer. They introduced masked tokens with various lengths based on the Poisson distribution, inspired by SpanBERT (Joshi et al., 2020), at multiple positions. BART is pretrained with large monolingual corpora (160 GB), including news, books, stories, and web-text domains. This model achieved strong results in several generation tasks; thus, it is regarded as a generic model.

They released pretrained models using English monolingual corpora for several tasks, including summarization, which we used for English GEC. Liu et al. (2020) proposed multilingual BART (mBART) for a machine translation task, which we used for GEC of several languages. The latter model was trained using monolingual corpora for 25 languages simultaneously. They used a special token for representing the language of a sentence. For example, they added `<de_DE>` and `<ru_RU>` into the initial token of the encoder and decoder for De–Ru translation. To fine-tune mBART for German, Czech, and Russian GEC, we set the target language for the special token referring to that language.

4 Experiment

4.1 Settings

Common Settings. As presented in Table 1, we used learner corpora, including BEA[2] (Bryant et al., 2019; Granger, 1998; Mizumoto et al., 2011; Tajiri et al., 2012; Yannakoudakis et al., 2011; Dahlmeier et al., 2013), JFLEG (Napoles et al., 2017), and CoNLL-14 (Ng et al., 2014) data for

[2] BEA corpus is made of several corpora. Details can be found in Bryant et al. (2019).

lang	Corpus	Train	Dev	Test
En	BEA	1,157,370	4,384	4,477
	JFLEG	-	-	747
	CoNLL-2014	-	-	1,312
De	Falko+MERLIN	19,237	2,503	2,337
Cz	AKCES-GEC	42,210	2,485	2,676
Ru	RULEC-GEC	4,980	2,500	5,000

Table 1: Data statistics.

English; Falko+MERLIN data (Boyd et al., 2014) for German; AKCES-GEC (Náplava and Straka, 2019) for Czech; and RULEC-GEC (Rozovskaya and Roth, 2019) for Russian.

Our models were fine-tuned using a single GPU (NVIDIA TITAN RTX), and our implementations were based on publicly available code[3]. We used the hyperparameters provided in some previous works (Lewis et al., 2020; Liu et al., 2020), unless otherwise noted.

The scores excluding the ensemble method were averaged in five fine-tuned experiments with random seeds.

English. Our setting for the English datasets was almost the same as that of Kiyono et al. (2019). We extracted the training data from BEA-train for English GEC. Similar to Kiyono et al. (2019), we did not use the unchanged sentences in the source and target sides; thus, the training data consisted of 561,525 sentences. We used BEA-dev to determine the best model.

We trained the BART-based models by using `bart.large`. This model was proposed for the summarization task, which required some constraints in inference to ensure appropriate outputs; however, we did not impose any constraints because our task was different. We applied byte pair encoding (BPE) (Sennrich et al., 2016) to the training data for the BART-based model by using the BPE model of Lewis et al. (2020).

We used the M^2 scorer (Dahlmeier and Ng, 2012) and GLEU (Napoles et al., 2015) for CoNLL-14 and JFLEG, respectively, and used the ERRANT scorer (Bryant et al., 2017) for BEA-test. We compared these scores with strong results (Kiyono et al., 2019; Kaneko et al., 2020).

German, Czech, and Russian. The dataset settings in this study were almost the same as those used by Náplava and Straka (2019) for each language. We used official training data and decided the best model by using the development data.

In addition, we trained the mBART-based models for German, Czech, and Russian GEC. We used `mbart.cc25` for the mBART-based models. For the mBART-based model, we followed Liu et al. (2020); we detokenized[4] the GEC training data for the mBART-based model and applied SentencePiece (Kudo and Richardson, 2018) with the SentencePiece model shared by Liu et al. (2020). Using this preprocessing, the input sentence may not represent grammatical information, compared with the sentence tokenized using a morphological analysis tool and subword tokenizer. However, what preprocessing is appropriate for GEC is beyond this paper's scope and will be treated as future work. For evaluation, we tokenized the outputs after recovering the subwords. Then, we used a spaCy-based[5] tokenizer for German[6] and Russian[7], and the MorphoDiTa tokenizer[8] for Czech.

Moreover, the M^2 scorer was used for each language. We compared these scores with the current SOTA results (Náplava and Straka, 2019).

4.2 Results

English. Table 2 presents the results of the English GEC task. When using a single model, the BART-based model is better than the model proposed by Kiyono et al. (2019), and the results are comparable to those reported by Kaneko et al. (2020) in terms of CoNLL-14 and BEA-test. Kiyono et al. (2019) and Kaneko et al. (2020) incorporated several techniques to improve the accuracy of GEC. To compare these models, we experimented with an ensemble of five models. Our ensemble model was slightly better than our single model, but worse than the ensemble models by Kiyono et al. (2019) and Kaneko et al. (2020). The BART-based model along with the ensemble model achieved results comparable to current strong results despite only requiring fine-tuning of the BART model. We believe that the reason for the ineffectiveness of the ensemble method is that the five models are not significantly different as the

[3] BART, mBART: https://github.com/pytorch/fairseq

[4] We used detokenizer.perl in the Moses script (Koehn et al., 2007).

[5] https://spacy.io

[6] We used the built-in de model.

[7] https://github.com/aatimofeev/spacy_russian_tokenizer

[8] https://github.com/ufal/morphodita

	CoNLL-14 (M^2)			JFLEG	BEA-test		
	P	R	$F_{0.5}$	GLEU	P	R	$F_{0.5}$
Kiyono et al. (2019)	67.9/73.3	44.1/44.2	61.3/64.7	59.7/61.2	65.5/74.7	59.4/56.7	64.2/70.2
Kaneko et al. (2020)	69.2/72.6	**45.6**/46.4	**62.6**/65.2	**61.3**/62.0	67.1/72.3	**60.1**/61.4	**65.6**/69.8
BART-based	**69.3**/69.9	45.0/45.1	**62.6**/63.0	57.3/57.2	**68.3**/68.8	57.1/57.1	**65.6**/66.1

Table 2: English GEC results. Left and right scores represent single and ensemble model results, respectively. Bold scores represent the best score in the single models, and underlined scores represent the best overall score.

		P	R	$F_{0.5}$
De	Náplava and Straka (2019)	78.21	59.94	73.31
	mBART-based	73.97	53.98	68.86
Cz	Náplava and Straka (2019)	83.75	68.48	80.17
	mBART-based	78.48	58.70	73.52
Ru	Náplava and Straka (2019)	63.26	27.50	50.20
	mBART-based	32.13	4.99	15.38
	with pseudo corpus	53.50	26.35	44.36

Table 3: German, Czech, and Russian GEC results. These models are not an ensemble of multiple models.

	Kaneko et al. (2020)			BART-based		
Error Type	P	R	$F_{0.5}$	P	R	$F_{0.5}$
PUNCT	74.1	52.7	68.5	79.2	59.0	**74.1**
DET	73.7	72.9	73.5	76.3	71.1	**75.2**
PREP	73.4	69.1	**72.5**	71.2	64.8	69.9
ORTH	86.9	62.9	**80.8**	84.2	52.9	75.3
SPELL	83.1	79.5	**82.3**	84.7	55.2	76.5

Table 4: BEA-test scores for the top five error types, except for OTHER. Kaneko et al. (2020) and BART-based are ensemble models. Bold scores represent the best score for each error type.

initial weights are the same as those of the BART model, and seeds only affect minor changes, such as training data order, and so on.

German, Czech, and Russian. Table 3 presents the results for German, Czech, and Russian GEC.

In the German GEC task, the mBART-based model achieves 4.45 $F_{0.5}$ points lower than the model by Náplava and Straka (2019). This may be because Náplava and Straka (2019) pretrains the GEC model with only the target language, whereas mBART is pretrained with 25 languages, resulting in the information of other languages being included as noise.

In the Czech GEC task, the mBART-based model achieves 6.65 $F_{0.5}$ points lower than the model by Náplava and Straka (2019). Similar to the case of the German GEC results, we suppose that mBART includes noisy information.

Considering Russian GEC, the mBART-based model shows much lower scores than Náplava and Straka (2019)'s model. This may be because the training data for Russian GEC are scarce compared to those of German or Czech. To investigate the effect of corpus size, we additionally trained the mBART model with a 10M pseudo corpus, using the method proposed by Grundkiewicz et al. (2019), and fine-tuned it with the learner corpus to compensate for the low-resource scenario. The results presented in Table 3 support our hypothesis.

5 Discussion

BART as a simple baseline model. According to the German and Czech GEC results, the mBART-based model, in which we only fine-tuned the pretrained mBART model, achieves comparable scores with SOTA models. In other words, mBART-based models are considered to show sufficiently high performance for several languages without using a pseudo corpus. These results indicate that the mBART-based model can be used as a simple GEC baseline for several languages.

Performance comparison for each error type. We compare the BART-based model with Kaneko et al. (2020)'s model for common error types using a generic pretrained model. Table 4 presents the results for the top five error types in BEA-test. According to these results, BART-based is superior to Kaneko et al. (2020) in PUNCT and DET errors; in particular, PUNCT is 5.6 $F_{0.5}$ points better. BART is pretrained to correct the shuffled and masked sequence, so that this model learns to place punctuation adequately. In contrast, Kaneko et al. (2020) uses an encoder that is not pretrained with correcting shuffled sequences.

Conversely, Kaneko et al. (2020) report better results for other errors, except for DET. Regarding ORTH and SPELL, their model is more than 5 $F_{0.5}$ points better than the BART-based one. It is difficult for the BART-based model to correct these errors because BART uses shuffled and masked sequences as noise in pretraining; not using character-level errors. Kaneko et al. (2020) introduce character errors into a pseudo corpus as task-oriented Enc–Dec pretraining; this is the reason why the BART-based model is inferior to Kaneko et al. (2020) in these errors.

6 Conclusion

We introduced a generic pretrained Enc–Dec model, BART, for GEC. The experimental results indicated that BART better initialized the Enc–Dec model parameters. The fine-tuned BART achieved remarkable results, which were comparable to the current strong results in English GEC. Indeed, the monolingual BART seems to be more effective for GEC than the model with a multilingual setting. However, although it is not as good as SOTA, fine-tuned mBART exhibited high performance in other languages. This implies that BART is a simple baseline model for pretraining GEC methods because it only requires fine-tuning as training.

Acknowledgements

We thank the anonymous reviewers for their insightful comments. This work has been partly supported by the programs of the Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (JSPS KAKENHI) Grant Numbers 19K12099 and 19KK0286.

References

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proc. of LREC*, pages 1281–1288.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proc. of ACL*, pages 793–805.

Christopher Bryant, Mariano Felice, ¥Oistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proc. of BEA*, pages 52–75.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proc. of NAACL-HLT*, pages 568–572.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proc. of BEA*, pages 22–31.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171–4186.

Sylviane Granger. 1998. The computer learner corpus: A versatile new source of data for SLA research. In Sylviane Granger, editor, *Learner English on Computer*, pages 3–18. Addison Wesley Longman.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proc. of BEA*, pages 252–263.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Can encoder-decoder models benefit from pre-trained language representation in grammatical error correction? In *Proc. of ACL*.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *Proc. of EMNLP-IJCNLP*, pages 1236–1242.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL Demo Sessions*, pages 177–180.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. of EMNLP: System Demonstrations*, pages 66–71.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xiongmin Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *ArXiv*, abs/2001.08210.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proc. of IJCNLP*, pages 147–155.

Jakub Náplava and Milan Straka. 2019. Grammatical error correction in low-resource scenarios. In *Proc. of W-NUT*, pages 346–356.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proc. of ACL-IJCNLP*, pages 588–593.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proc. of EACL*, pages 229–234.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proc. of CoNLL Shared Task*, pages 1–14.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proc. of BEA*, pages 163–170.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of ACL*, pages 1715–1725.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proc. of ACL*, pages 198–202.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*, pages 5998–6008.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proc. of ACL-HLT*, pages 180–189.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proc. of NAACL-HLT*, pages 156–165.