Artificial Error Generation with Fluency Filtering

Mengyang Qiu[†‡] **Jungyeul Park**[†]
[†] Department of Linguistics
[‡] Department of Communicative Disorders and Sciences
State University of New York at Buffalo
{mengyang,jungyeul}@buffalo.edu

Abstract

The quantity and quality of training data plays a crucial role in grammatical error correction (GEC). However, due to the fact that obtaining human-annotated GEC data is both time-consuming and expensive, several studies have focused on generating artificial error sentences to boost training data for grammatical error correction, and shown significantly better performance. The present study explores how fluency filtering can affect the quality of artificial errors. By comparing artificial data filtered by different levels of fluency, we find that artificial error sentences with low fluency can greatly facilitate error correction, while high fluency errors introduce more noise.

1 Introduction

Grammatical Error Correction (GEC), a NLP task of automatically detecting and correcting grammatical errors in text, has received much attention in the past few years, because of an ever-growing demand for reliable and quick feedback to facilitate the progress of English learners. In a typical GEC task, an error sentence such as *I* ***follows*** *his advice* needs to be corrected to a grammatical sentence *I* ***follow*** *his advice*, while a grammatical sentence *She follows his advice* should output the same sentence without any modification. Currently, neural machine translation (NMT) systems using sequence-to-sequence (seq2seq) learning (Sutskever et al., 2014) that “translate” incorrect sentences into correct ones, have shown to be promising in grammatical error correction, and several recent NMT approaches have obtained the state-of-the-art results in GEC (e.g., Chollampatt and Ng, 2018; Ge et al., 2018; Zhao et al., 2019).

While designing a GEC-oriented seq2seq architecture is one important aspect to achieve high performance in grammatical error correction, the quantity and quality of data also plays a crucial role in the NMT approach to GEC, as NMT parameters cannot learn and generalize well with limited training data. Due to the fact that obtaining human-annotated GEC data is both time-consuming and expensive, several studies have focused on generating artificial error sentences to boost training data for grammatical error correction. One main approach is to extract errors and their surrounding context (the context window approach) from available annotated data, and then apply the errors to error-free sentences naively or probabilistically (Yuan and Felice, 2013; Felice, 2016). The other approach uses machine back-translation, which switches the source-target sentence pairs in GEC and learns to ”translate” correct sentences into their incorrect counterparts (Kasewa et al., 2018). While the first approach may not generalize well to unseen errors, and the second one may have no control over what kind of error is produced, artificial error sentences generated from both approaches contribute to better performance in grammatical error correction.

In this paper, we do not focus on which approach is superior in artificial error generation. Rather, given that both approaches can generate multiple error candidates for each correct sentence, we investigate how to select the best ones that can boost GEC performance the most. Although previous studies have shown that artificial errors that match the real error distributions tend to generate better results (Felice, 2016; Xie et al., 2018), we propose an alternative framework that incorporates fluency filtering based on language models. We evaluate four strategies of artificial error selection using different fluency ranges (from lowest to highest) on the recent W&I+LOCNESS test set. Our results show that three of the four strategies lead to evident improvement over the original baseline, which is in line with previous findings that in general GEC benefits from artifi-

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 87–91
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

cial error data. The model trained with artificial error sentences with the lowest fluency obtains the highest recall among the four settings, while the one trained with error sentences with the median fluency achieves the highest performance in terms of $F_{0.5}$, with an absolute increase of 5.06% over the baseline model.

2 Related Work

Our work mainly builds on the context window approach to artificial error generation. In this approach, all the possible error fragments (errors and their surrounding context) and their corresponding correct fragments are first extracted from GEC annotated corpora. For example, *I **follows** his* and *I **follow** his* are the fragments extracted from the example sentences in the first paragraph. With these correct-incorrect fragments, for each error-free sentence, if we find the same correct fragment in the sentence, we can inject errors by replacing that fragment with the incorrect one. Felice (2016) has shown that a context window size of one, that is, one token before and after the error words or phrases, is able to generate a decent amount of error sentences while maintaining the plausibility of the errors. Thus, the current study also adopts this context window size in extracting fragments.

The current work is also inspired by the fluency boost learning proposed in Ge et al. (2018). In their study, sentence fluency is defined as the inverse of the sentence's cross entropy. During fluency boost training, the fluency of candidate sentences generated by their GEC seq2seq model is monitored. Candidate sentences with less than perfect fluency compared to the correct ones are appended as additional error-contained data for subsequent training. Fluency is also used during multi-round GEC inference, in that inference continues as long as the fluency of the output sentences keeps improving. The present study uses fluency measure in an opposite way. We examine how the decrease of fluency in artificial error sentences influences the performance of grammatical error correction.

3 Proposed Methods

To filter candidate error sentences based on fluency, our first step is to generate all the candidate sentences. With correct-incorrect fragment pairs extracted from GEC annotated corpora, we replace all correct fragments found in each error-free sentence with their incorrect counterparts exhaustively. Unlike a method described in Felice (2016) that multiple errors can apply to one sentence at the same time, we only allow one error at a time. Table 1 shows an example of an error-free sentence and the candidate sentences after applying all the possible error replacements. There is only one error in each candidate sentence, and the same position in the correct sentence can have multiple different replacements (*e.g.*, *effects* → *impacts*|*effect*|*dealing*). We then calculate the fluency score of each candidate sentence and select the ones with the highest fluency, lowest fluency and median fluency. Fluency score is measured by sentence perplexity, the inverse probability of the sentence based on a language model, normalized by the number of words in that sentence. A sentence's fluency score is negatively related to its perplexity. Our prediction is that low sentence fluency (high perplexity) can facilitate error detection and correction by maximizing and highlighting the difference between correct and incorrect sentences. Conversely, artificial error sentences of high fluency can be confusing to the model as the difference between correct and incorrect sentences may be subtle.

4 Experiments and Results

4.1 Dataset and evaluation

We used the four datasets — FCE, NUCLE, W&I+LOCNESS and Lang-8 — provided in the BEA 2019 Shared Task on GEC[1] as the training data for our baseline model (in total about 1.1M sentence pairs). Table 2 shows the summary of the four datasets. There are slightly over half a million error-contained sentences in these datasets, where we extracted 1.3M correct-incorrect fragments. We applied our artificial error injection procedure to the remaining 0.6M error-free sentences, and over 0.4M of them received replacements. We trained a 3-gram language model on all the correct-side sentences using KenLM (Heafield, 2011). The language model was used to calculate perplexity of artificial error sentences. From the 0.4M sentences with error injections, we created four different artificial datasets: one with the highest fluency error sentences among the candidates of each correct sentence, one with the lowest, one with the median, and the last one was

[1] `https://www.cl.cam.ac.uk/research/nl/bea2019st/`

	Sentence	Fluency
Correct	the **effects** of **the use** of biometric identification are obvious .	
Candidates:	the effects of **the used** of biometric identification are obvious .	
	the effects of **use** of biometric identification are obvious .	Median
	the effects of **the using** of biometric identification are obvious .	
	the **impacts** of the use of biometric identification are obvious .	
	the **effect** of the use of biometric identification are obvious .	Highest
	...	
	the **dealing** of the use of biometric identification are obvious .	Lowest

Table 1: An example of an error-free sentence and its error injected candidate sentences with three levels of fluency.

Corpus	# Sent Pairs
FCE (Train)	28,346
NUCLE	57,113
W&I+LOCNESS (Train)	34,304
LANG-8	1,037,561
Total	1,157,324
Correct	601,958
Error Injection to Correct	444,521

Table 2: Summary of training data.

randomly selected. Each version was then combined with the original error-contained sentences and the remaining unchanged correct sentences so that all these settings had the same number of sentence pairs as in our baseline model (1.1M). The goal of the experiment was to compare the GEC performance trained with these four settings to the baseline. The W&I+LOCNESS development set of 4,382 sentences was used as validation, and the W&I+LOCNESS test set of 4,477 sentences as evaluation[2]. All these settings were run for three times. Performance was evaluated in terms of precision, recall and $F_{0.5}$ using ERRANT (Bryant et al., 2017).

4.2 Experimental settings

We used the 7-layer convolutional seq2seq model[3] proposed in Chollampatt and Ng (2018) for grammatical error correction with minimal modification. The only difference to Chollampatt and Ng (2018) is that the word embedding dimensions in both encoders and decoders were set to 300 rather than 500, and the word embeddings were trained separately using the error and correct side training data instead of external corpora. Other parameters were set as recommended in Chollampatt and Ng (2018), including the top 30K BPE tokens as the vocabularies of input and output, 1,024 × 3 hidden layers in the encoders and decoders, Nesterov Accelerated Gradient as the optimizer with a momentum of 0.99, dropout rate of 0.2, initial learning rate of 0.25, and minimum learning rate of 10^{-4}. A beam size of 10 was used during model inference. No spell checker was incorporated in the present study, either as pre-processing or post-processing.

4.3 Experimental results

Table 3 shows the results for our baseline and models trained with different fluency-filtered artificial errors. The model trained on the baseline data, which include 0.6M correct sentence pairs, performs the worst in terms of recall (18.85%), because the large proportion of the same sentences makes the model too conservative to make corrections. Indeed, true positive for the baseline model is only 749, which is about half of that in the lowest fluency condition. All the four models with artificial errors obtain higher recall (over 26%), but at the expense of precision. The model with error sentences that have the highest fluency among candidate sentences, in particular, drops over 15% in precision compared to the baseline, making it the worst model in terms of $F_{0.5}$ (42.86%). Error sentences with the lowest fluency lead to the highest recall (32.96%) and second highest $F_{0.5}$ (48.68%) among all the models, while the model in the median fluency condition achieves a good balance between precision drop and recall gain, resulting in the highest $F_{0.5}$ (49.03%).

[2] https://competitions.codalab.org/competitions/21922

[3] https://github.com/pytorch/fairseq

	Prec.	Recall	$F_{0.5}$
Original (Baseline)	**65.93**	18.85	43.97
Random	55.67	27.61	46.26
Highest	50.44	26.77	42.86
Median	**57.69**	30.64	**49.03**
Lowest	55.27	**32.96**	48.68

Table 3: Performance of multi-layer CNNs for GEC on W&I+LOCNESS test set with different error data from different fluency filtering.

5 Conclusions and Future Work

The goal of the current study was to explore how the fluency of artificial error sentences can affect the performance of grammatical error correction. We greedily generated all possible error sentences using the similar context window approach as in Felice (2016), and then selected among candidate sentences based on fluency score (sentence perplexity). As predicted, the model trained with artificial error sentences of highest fluency performed even worse than the baseline model with a large proportion of correct sentence pairs. Models in both lowest and median fluency conditions performed significantly better than the other three models. The former one achieved the highest recall, while the latter one was more balanced with the highest $F_{0.5}$. These results indicate that fluency filtering can be used a means to select high-quality artificial error sentences for grammatical error detection and correction.

Although the present study just focused on fluency and ignored error probability, the two factors are not mutually exclusive. Rather, combining the two approaches may generate even better artificial errors. Additionally, fluency filtering is not restricted to the context window approach to error generation, it can be part of the machine back-translation approach to help select among the N best translations.

One limitation of the current study is that we only generated one error at a time for each sentence. In the training data, the 0.5M error sentences contain 1.3M errors, which means that on average each sentence has 2.4 errors. Our next step is to explore using fluency filtering to ensure the quality of artificial multi-error sentences and to see if this can boost GEC performance even further.

References

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Shamil Chollampatt and Hwee Tou Ng. 2018. A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction. In *Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, Louisiana.

Mariano Felice. 2016. Artificial error generation for translation-based grammatical error correction. Technical Report UCAM-CL-TR-895, University of Cambridge, Computer Laboratory.

Tao Ge, Furu Wei, and Ming Zhou. 2018. Fluency Boost Learning and Inference for Neural Grammatical Error Correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1055–1065, Melbourne, Australia. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. 2018. Wronging a Right: Generating Better Errors to Improve Grammatical Error Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4977–4983, Brussels, Belgium. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. 2018. Noising and Denoising Natural Language: Diverse Backtranslation for Grammar Correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 619–628, New Orleans, Louisiana. Association for Computational Linguistics.

Zheng Yuan and Mariano Felice. 2013. Constrained grammatical error correction using statistical machine translation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 52–61, Sofia, Bulgaria. Association for Computational Linguistics.

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 *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota. Association for Computational Linguistics.