Artificial Error Generation with Fluency Filtering

Mengyang Qiu[†‡] Jungyeul Park[†]

[†] Department of Linguistics

[‡] Department of Communicative Disorders and Sciences

{mengyang,jungyeul}@buffalo.edu

University at Buffalo
The State University of New York

Abstract

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. The current study investigates how to select the best artificial error sentences among candidate sentences 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 significant improvement over the original baseline, which is in line with previous findings that in general GEC benefits from artificial 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 $F_{0.5}$ score, with an absolute increase of 5.06% over the baseline model.

Proposed Methods

- The current study examines how the decrease of fluency (sentence perplexity) in artificial error sentences influences the performance of grammatical error correction.
- 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.
- For each sentence, only one error is allowed at a time. The same position in the correct sentence can have multiple different replacements.
- We then calculate the fluency score of each candidate sentence and select the ones with the highest fluency, lowest fluency and median fluency.
- Our prediction is that low sentence fluency (high perplexity) can facilitate error correction by maximizing the difference between correct and incorrect sentences. Conversely, high fluency error sentences can be confusing to the model as the difference may be subtle.

	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

Experimental Settings

- We apply our artificial error generation procedures to the 0.6M error-free sentences in the GEC corpora. These error-injected sentences, together with the original 0.5M error sentences, are our experimental training data. The original 1.1M sentences without error injection are used as our baseline.
- We create four different artificial datasets: error sentences with highest fluency, with lowest fluency, with median fluency, and randomly picked error sentences.

NMT Training

- 7-layer convolutional seq2seq model proposed in Chollampatt and Ng (2018)
- Top 30K BPE tokens as vocabularies
- Word embedding dimensions: 300
- 1,024(hidden size) $\times$ 3(window size) in the hidden layers
- Nesterov Accelerated Gradient as the optimizer with a momentum of 0.99, dropout rate of 0.2 and an adaptive learning rate (initially 0.25, minimum 10^{-4})

Summary of Training Data

Corpora	# Sent Pairs
FCE	28,346
NUCLE	57,113
W&I+LOCNESS	34,304
LANG-8	1,037,561
Total	1,157,324
Error-free	601,958

Results and Conclusion

- The baseline data with 6M unchanged 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.
- All the experimental models with artificial errors obtain higher recall (over 26%), but at the expense of precision. The highest fluency condition, 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%).
- One limitation of the current study is that we only generate one error for each sentence. In the training data, the 0.5M error sentences contain 1.3M errors (on average 2.4 errors per sentence). Our next step is to explore generating artificial multi-error sentences and to see if this can boost GEC performance even further.

	Prec.	Recall	$F_{0.5}$
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

Selected 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 Net-work 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.