Automatic Grammatical Error Correction for Sequence-to-sequence Text Generation: An Empirical Study

Tao Ge **Xingxing Zhang** **Furu Wei** **Ming Zhou**
Microsoft Research Asia
{tage, xizhang, fuwei, mingzhou}@microsoft.com

Abstract

Sequence-to-sequence (seq2seq) models have achieved tremendous success in text generation tasks. However, there is no guarantee that they can always generate sentences without grammatical errors. In this paper, we present a preliminary empirical study on whether and how much automatic grammatical error correction can help improve seq2seq text generation. We conduct experiments across various seq2seq text generation tasks including machine translation, formality style transfer, sentence compression and simplification. Experiments show the state-of-the-art grammatical error correction system can improve the grammaticality of generated text and can bring task-oriented improvements in the tasks where target sentences are in a formal style.

1 Introduction

Sequence-to-sequence (seq2seq) text generation (Cho et al., 2014; Sutskever et al., 2014) has attracted growing attention in natural language processing (NLP). Despite various advantages of seq2seq models, they tend to have a weakness: there is no guarantee that they can always generate sentences without grammatical errors. Table 1 shows examples generated by seq2seq models in various tasks *with* grammatical errors.

One valid solution to this challenge is conducting grammatical error correction (GEC) for machine generated sentences. Recent GEC systems (Chollampatt and Ng, 2018; Junczys-Dowmunt et al., 2018; Grundkiewicz and Junczys-Dowmunt, 2018; Ge et al., 2018a,b) can achieve human-level performance in GEC benchmarks. We are curious whether they can help improve seq2seq based natural language generation (NLG) models. We therefore propose an empirical study on GEC post editing for various text generation tasks (i.e., machine translation, style transfer, sentence compression and simplification) using both automatic and human evaluation methods. Experimental results demonstrate that a state-of-the-art GEC system is helpful for improving the grammaticality of generated text and that it can bring task-oriented improvements in the tasks where target sentences are in a formal style.

Tasks	Examples
Machine Translation	Das Team-Ereignis ist immer am besten. → The team event is always **(the)** best.
Style Transfer	=) who do u thinks better? → Who do you think **(is)** better?
Sentence Compression	Mickey Rooney died yesterday age 93 at his home in Studio City, California... → Mickey Rooney died yesterday **(at)** age 93.

Table 1: Seq2seq model outputs for German-English translation, formality style transfer and sentence compression. The texts in round brackets are edits by GEC.

The contributions of this paper are twofold:

- We present an empirical study on GEC post editing for seq2seq text generation. To the best of our knowledge, it is the first work to study improving seq2seq based NLG models using GEC.
- We show some interesting results by thoroughly comparing and analyzing GEC post editing for various seq2seq text generation tasks, shedding light on the potential of GEC for NLG.

2 Background

2.1 Sequence-to-sequence Text Generation

The sequence-to-sequence (seq2seq) framework has been proven to be successful for many NLP tasks. Given a source sentence $\boldsymbol{x^s}$, a seq2seq model learns to predict its target sentence $\boldsymbol{x^t}$. It usually has an encoder to learn the representation of $\boldsymbol{x^s}$ and a decoder to generate $\boldsymbol{x^t}$ based on the encoded representation of $\boldsymbol{x^s}$. The model is

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6059–6064
Florence, Italy, July 28 - August 2, 2019.

usually trained by minimizing the negative log-likelihood of the training source-target sentence pairs. During inference, an output sequence $\boldsymbol{x^o}$ is generated (one token at a time) with beam search by maximizing $P_{\Theta}(\boldsymbol{x^o}|\boldsymbol{x^s})$.

2.2 Automatic Grammatical Error Correction

Most recent GEC systems are based on the seq2seq framework and are trained with error-corrected sentence pairs. Due to massive training data, the state-of-the-art GEC system (Grundkiewicz and Junczys-Dowmunt, 2018; Ge et al., 2018b) can achieve human-level performance in GEC benchmarks and be practically used for correcting grammatical errors.

3 Experiments and Evaluation

We use the state-of-art GEC system (Ge et al., 2018b) as our GEC model which is a 7-layer convolutional seq2seq model trained with a fluency boost learning strategy on both original GEC training data and augmented fluency boost sentence pairs. We use the GEC model to do post editing for sentences decoded by a seq2seq model to test if GEC improves the results. We choose machine translation, style transfer, sentence compression and simplification as typical seq2seq text generation tasks.

Due to the page limit, the detailed configuration of the models we implemented in this section are put in the supplementary notes.

3.1 Machine translation

We take Machine translation (MT) as the main task to study whether GEC helps improve translation quality. We conduct experiments by using GEC to edit the results of the state-of-the-art neural machine translation (NMT) system (Google Translate) on the French-English (FR-EN) in WMT14, German-English (DE-EN) and Chinese-English (ZH-EN) news test sets in WMT17.

Table 2 shows BLEU with/without post-editing by the GEC system. Although GEC post-editing does not improve BLEU much, when we look into the results by analyzing the sentences edited by GEC, we observe only a small proportion of sentences are modified by the GEC system – approximately 5% in FR-EN and DE-EN, while 10% in ZH-EN test sets. The sentence-level BLEU of around 50% of the edited sentences are improved,

	NMT	NMT+GEC	#edited
FR-EN	**38.70**	38.69 $_{(-0.01)}$	131 (63↑ 68↓) out of 3,003
DE-EN	35.45	**35.48** $_{(+0.03)}$	141 (65↑ 76↓) out of 3,004
ZH-EN	28.85	**28.96** $_{(+0.11)}$	271 (148↑ 123↓) out of 2,001

Table 2: BLEU with/without post editing by GEC. **#edited** shows the number of sentences modified by GEC, where ↑ and ↓ indicate the number of sentences whose BLEU improves or decreases.

	MT	MT+GEC
Unsupervised SMT	27.09	**27.33** $_{(+0.24)}$
Unsupervised NMT	28.30	**28.52** $_{(+0.22)}$
Google Translate	38.70	38.69 $_{(-0.01)}$

Table 3: BLEU of the unsupervised SMT and NMT systems in the WMT14 FR-EN test set.

while the remaining suffer a BLEU decrease.

To understand the reasons for the BLEU changes, we manually check each sentence edited by GEC in WMT14 FR-EN dataset and show the results in Table 4. The main reason (90.5% cases) for a BLEU improvement is that GEC corrects errors in NMT's results and improves the translation quality. In contrast, the reasons why BLEU decreases are various. First, the correction of grammatical errors by GEC may decrease BLEU though it improves the sentence's grammaticality, as shown in Table 4. Second, the GEC system is not perfect: it sometimes edits a sentence without grammatical errors. Even though such edits usually bring no adverse effects, it is likely to decrease BLEU. Last, we find reference sentences occasionally have grammatical errors, as *Reference Error* in Table 4 shows. When GEC fixes the errors in such cases, BLEU decreases.

Moreover, we test the effects of GEC on MT in a low resource setting. We use the state-of-the-art unsupervised SMT and NMT model in Ren et al. (2019) and use the GEC system to edit their results. According to the results shown in Table 3, the unsupervised MT systems benefit more from GEC than the state-of-the-art supervised NMT (i.e, Google translate) because they are more likely to generate sentences that are not fluent than the supervised MT models, which can be addressed by GEC.

We also conduct experiments on the WMT17 Automatic Post-Editing (APE) task. However, we observe a large number of grammatical errors in the references which make the automatic evaluation less reliable. We include the results in the supplementary notes due to the page limit.

BLEU change	Reasons	Examples
BLEU↑ (63)	Correction (90.5%)	NMT: They know their business better than anyone. (76.7) GEC: They know their business better than anyone else. (100) REF: They know their business better than anyone else.
	Accidental (9.5%)	NMT: But this pacified identity only had a time. (12.1) GEC: But this pacified identity only had time. (12.2) REF: Yet, this pacified identity has had its day.
BLEU↓ (68)	Correction (52.9%)	NMT: It's good child, it's cool. (51.3) GEC: It's a good child, it's cool. (45.2) REF: It's relaxed, it's cool.
	GEC Error (30.9%)	NMT: At the piano, dancers take turns to play the scores. (100) GEC: At the piano, dancers take turns playing the scores. (64.1) REF: At the piano, dancers take turns to play the scores.
	Reference Error (16.2%)	NMT: FAA may lift ban on certain electronic devices during take-off and landing (46.6) GEC: FAA may lift a ban on certain electronic devices during take-off and landing (16.3) REF: FAA may lift ban on some electronic devices during takeoff and landing

Table 4: Reasons for BLEU changes in WMT14 FR-EN dataset. The numbers in the round brackets following example sentences are sentence-level BLEU.

	Informal→Formal		Formal→Informal	
	BLEU	Acc	BLEU	Acc
Transformer	73.79	83.0	38.49	68.7
Transformer+GEC	**74.84**	**84.2**	**38.85**	47.1
State-of-the-art	75.37	-	39.09	-

Table 5: Results for GEC post-editing on formality style transfer on the GYAFC test set in "Family & Relationships" domain, containing about 1,000 sentences. Acc is evaluated with the help of a CNN model for style classification. The state-of-the-art (Niu et al., 2018) is an ensemble model trained with additional data.

3.2 Formality style transfer

In addition to MT, we test GEC on the text style transfer task. We study formality style transfer which transfers an informal (formal) sentence to a formal (informal) style and choose GYAFC corpus (Rao and Tetreault, 2018) as our testbed. We use a 2-layer transformer model as our base model and train a model with approximately 100K parallel sentences in the GYAFC corpus for informal→informal and formal→informal respectively. We use the GEC model to edit the base models' outputs, and show the result in Table 5.

While GEC improves BLEU in both transfer directions, we observe differences when we look into style accuracy. For Informal→Formal transfer, accuracy is improved (83.0% → 84.2%) after GEC post editing; while for Formal→Informal transfer, it decreases (68.7% → 47.1%) because grammaticality improvements by GEC may make a sentence become less like an informal sentence.

3.3 Sentence compression and simplification

We also test effects of GEC post-editing on sentence compression and simplification. For sentence compression, following Filippova et al. (2015), we train a 2-layer LSTM seq2seq model, which generates a 0/1 sequence to indicate whether to delete a word, as our base model and test on Google's sentence compression dataset[1] (GoogComp). For sentence simplification, we use the state-of-the-art deep reinforcement model DRESS (Zhang and Lapata, 2017) as our base model and test on Newsela text simplification dataset.

Table 6 shows the results for the effects of GEC on sentence compression and simplification. For sentence compression, BLEU decreases from 60.38 to 58.77 after GEC post editing. We manually analyze the results and find there are many grammatical errors in the reference sentences. This is not surprising, since the reference sentences are constructed with an automatic approach (Filippova and Altun, 2013). The grammatical errors in the references affect the BLEU evaluation and make it less reliable.

The BLEU decrease is also observed in sentence simplification task but for a different reason. In the Newsela dataset, the reference sentences are written by humans and therefore have much fewer grammatical errors compared to GoogComp. In contrast to sentence compression where reference errors are the main reason for the BLEU decrease, the BLEU decrease in sentence simplification usually happens in the cases where the correction of grammatical errors reduces the sentence's n-gram overlap with the reference sentence, as shown in Table 6 (similar to the phenomenon observed in the experiments for MT; see Table 4). In addition, GEC errors and occasional errors in reference sen-

[1] https://github.com/google-research-datasets/sentence-compression

Tasks	#edited	BLEU	BLEU change	Reasons	Examples
Sentence Compression	110	60.38 ↓ 58.77	10↑	Accidental (100%)	Base: Domestic flights were cancelled Sunday. (9.4) GEC: Domestic flights were cancelled on Sunday. (9.9) REF: Several domestic flights were cancelled due to the bad weather.
			100↓	Reference Error (45.0%)	Base: An tanker caught fire in a garage. (100) GEC: A tanker caught fire in a garage. (84.1) REF: An tanker caught fire in a garage.
				Correction (37.0%)	Base: A undersea earthquake shook eastern Indonesia. (72.9) GEC: An undersea earthquake shook eastern Indonesia. (70.1) REF: A strong undersea earthquake shook eastern Indonesia.
				GEC Error (18.0%)	Base: Nine persons were arrested over the weekend. (25.1) GEC: Nine people were arrested over the weekend. (4.8) REF: Nine persons were arrested after a series of drug finds.
Sentence Simplification	96	22.64 ↓ 22.54	41↑	Correction (51.2%)	Base: She also speak to younger women who are interested in science and math. (77.4) GEC: She also speaks to younger women who are interested in science and math. (85.6) REF: She speaks to younger women who are interested in science and math.
				Accidental (48.8%)	Base: For mining, there's the International Seabed Authority. (11.9) GEC: For mining, there is the International Seabed Authority. (12.4) REF: The International Seabed Authority is for mining.
			55↓	Correction (58.2%)	Base: The rocks moves forward for a few days. (1.3) GEC: The rocks move forward for a few days. (1.2) REF: The lava moves for a few days, then stops for weeks before starting again.
				GEC Error (36.4%)	Base: In 2010, a group of chimpanzees was sent from the Netherlands to a zoo in Scotland. (51.7) GEC: In 2010, a group of chimpanzees were sent from the Netherlands to a zoo in Scotland. (45.0) REF: In 2010, a group of chimpanzees was taken from a zoo in the Netherlands.
				Reference Error (5.5%)	Base: Richie wrote the winning word "magician." (35.5) GEC: Richie wrote the winning word "magician". (7.9) REF: The winning word was "magician."

Table 6: Results for sentence compression and sentence simplification. As in Table 4, the numbers in the round brackets following the example sentences are sentence-level BLEU.

tences lead to a decrease of BLEU after GEC post editing.

3.4 Human Evaluation

In addition to automatic evaluation (e.g., BLEU), we present human evaluation results for GEC post editing on the tasks. The evaluation includes two aspects: First, we evaluate how much helpful GEC is for improving the grammaticality of sentences generated by the seq2seq models, which is independent to a specific task; Second, we evaluate if GEC's edits bring task-oriented improvements. The evaluation is done by a human judge through comparing the results with/without GEC's edits.

Table 7 shows the human evaluation results. For most sentences edited by GEC, their grammaticality is improved; while the bad cases are only in a small proportion ($\leq$10%) in all the six tasks. In contrast, the task-oriented improvements vary across the tasks. For example, for Informal→Formal style transfer, GEC performs well because most of its edits improve the sentences' grammaticality and make the sentences become more formal; in contrast, for Formal→Informal style transfer, GEC improves sentences' grammaticality but affects their styles, making them become less informal.

Moreover, it is observed that GEC is more beneficial to the seq2seq models trained in a low resource setting, by comparing the results of supervised and unsupervised MT, which is consistent with results in Table 3. For sentence compression and simplification, many grammatical improvements do not bring task-oriented improvements. The reason is that the parts GEC edits are not the content that should be kept in the results. Also, it is notable that except for Formal→Informal style transfer whose target sentences should be in an informal style, GEC brings much more improvements than adverse effects on the tasks, demonstrating the potential of GEC for NLG.

4 Related Work and Discussion

The most related work to ours is the automatic post editing (APE) (Bojar et al., 2016) which has been extensively studied for MT (e.g., (Pal et al., 2016, 2017; Chatterjee et al., 2017; Hokamp, 2017; Tan et al., 2017)) in the past few years. These APE approaches are usually trained with source language input data, target language MT output and target language post editing (PE) data. Although these APE models and systems have proven to be successful in improving MT results, they are task-specific and cannot be used for other NLG tasks.

In contrast, we propose a general post editing approach by applying the current state-of-the-art GEC system to editing the outputs of NLG systems. To the best of our knowledge, this is the first attempt to explore improving seq2seq based NLG models with a state-of-the-art neural GEC system despite some early studies on post-processing SMT outputs using a (mainly rule-based) grammar checker (Stymne and Ahrenberg, 2010). Experiments show GEC post editing can effectively improve the grammaticality of generated text and lead to a task-oriented improvement in the NLG

Tasks	#edited / #all	Grammaticality			Task-oriented		
		↑	↓	→	↑	↓	→
Supervised FR-EN NMT	131 / 3,003	**79%**	10%	11%	**63%**	10%	27%
Unsupervised FR-EN NMT	474 / 3,003	**85%**	4%	11%	**80%**	4%	16%
Informal→Formal	143 / 1,332	**74%**	6%	20%	**61%**	6%	33%
Formal→Informal	259 / 1,019	**91%**	2%	7%	4%	**79%**	17%
Sentence compression	110 / 2,000	**75%**	10%	15%	44%	13%	44%
Sentence simplification	96 / 1,077	**79%**	9%	12%	**47%**	12%	41%

Table 7: Human evaluation results for the sentences edited by GEC. ↑, ↓ and → denote GEC makes a sentence better, worse and neither better nor worse. The percentages are the proportion of the corresponding cases.

tasks where target sentences are in a formal style, especially in a low-resource setting.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. Specially, we thank Shujie Liu for the discussion and constructive suggestions to this paper. We also thank Shuo Ren, Shuangzhi Wu, Zhirui Zhang and Yi Zhang for their help with the evaluation in the machine translation and formality style transfer task.

References

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *ACL 2016 FIRST CONFERENCE ON MACHINE TRANSLATION (WMT16)*, pages 131–198. The Association for Computational Linguistics.

Rajen Chatterjee, M Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017. Multi-source neural automatic post-editing: Fbks participation in the wmt 2017 ape shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 630–638.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. *arXiv preprint arXiv:1801.08831*.

Katja Filippova, Enrique Alfonseca, Carlos A Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368.

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491.

Tao Ge, Furu Wei, and Ming Zhou. 2018a. 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)*, volume 1, pages 1055–1065.

Tao Ge, Furu Wei, and Ming Zhou. 2018b. Reaching human-level performance in automatic grammatical error correction: An empirical study. *arXiv preprint arXiv:1807.01270*.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2018. Near human-level performance in grammatical error correction with hybrid machine translation. *arXiv preprint arXiv:1804.05945*.

Chris Hokamp. 2017. Ensembling factored neural machine translation models for automatic post-editing and quality estimation. *arXiv preprint arXiv:1706.05083*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. *arXiv preprint arXiv:1804.05940*.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. *arXiv preprint arXiv:1806.04357*.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 281–286.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, Qun Liu, and Josef van Genabith. 2017. Neural automatic post-editing using prior alignment and reranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 349–355.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 129–140.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. *arXiv preprint arXiv:1901.04112*.

Sara Stymne and Lars Ahrenberg. 2010. Using a grammar checker for evaluation and postprocessing of statistical machine translation. In *LREC*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Yiming Tan, Zhiming Chen, Liu Huang, Lilin Zhang, Maoxi Li, and Mingwen Wang. 2017. Neural post-editing based on quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 655–660.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.