Cross-Corpora Evaluation and Analysis of Grammatical Error Correction Models — Is Single-Corpus Evaluation Enough?

Masato Mita[1,2]**,Tomoya Mizumoto**[1]**,Masahiro Kaneko**[3,1]**,Ryo Nagata**[4,1]**,Kentaro Inui**[2,1]
[1]RIKEN AIP, [2]Tohoku University, [3]Tokyo Metropolitan University, [4]Konan University
{masato.mita, tomoya.mizumoto}@riken.jp,
kaneko-masahiro@ed.tmu.ac.jp,nagata-naacl@ml.hyogo-u.ac.jp.,
inui@ecei.tohoku.ac.jp

Abstract

This study explores the necessity of performing cross-corpora evaluation for grammatical error correction (GEC) models. GEC models have been previously evaluated based on a single commonly applied corpus: the CoNLL-2014 benchmark. However, the evaluation remains incomplete because the task difficulty varies depending on the test corpus and conditions such as the proficiency levels of the writers and essay topics. To overcome this limitation, we evaluate the performance of several GEC models, including NMT-based (LSTM, CNN, and transformer) and an SMT-based model, against various learner corpora (CoNLL-2013, CoNLL-2014, FCE, JFLEG, ICNALE, and KJ). Evaluation results reveal that the models' rankings considerably vary depending on the corpus, indicating that single-corpus evaluation is insufficient for GEC models.

1 Introduction

Grammatical error correction (GEC) is the task of correcting various grammatical errors in a given text, which is typically written by non-native speakers. Previous studies focused on typical errors such as those in the use of articles (Han et al., 2006), prepositions (Felice and Pulman, 2008), and noun numbers (Nagata et al., 2006). Machine translation approaches are being presently applied for GEC (Junczys-Dowmunt et al., 2018; Chollampatt and Ng, 2018; Ge et al., 2018; Junczys-Dowmunt and Grundkiewicz, 2016). In these approaches, GEC is treated as a translation problem from the erroneous text to the correct text (Mizumoto et al., 2012; Felice et al., 2014; Junczys-Dowmunt and Grundkiewicz, 2014).

However, the evaluation of GEC performance is unfortunately not complete because researchers tend to evaluate their models on a single corpus. The CoNLL-2014 shared task dataset (Ng et al., 2014) has been recently used for such evaluation.

Single-corpus evaluation may be insufficient in cases wherein a GEC model generally aims to robustly correct grammatical errors in *any* written text partly because the task difficulty varies depending on proficiency levels and essay topics. Although a model outperforms a baseline in one corpus, the model in another corpus may perform better, leading to different conclusions from what we know. This study explores the necessity of performing cross-corpora evaluation for GEC models. The performance of four recent models, namely three neural machine translation (NMT)-based models (LSTM, CNN, and transformer) and a statistical machine translation (SMT)-based model is evaluated against six learner corpora (CoNLL-2014, CoNLL-2013 (Ng et al., 2013), FCE (Yannakoudakis et al., 2011), JFLEG (Napoles et al., 2017), KJ (Nagata et al., 2011), and ICNLAE (Ishikawa, 2013)). Evaluation results show that the models' rankings considerably vary depending on the corpus. Empirical results reveal that models must be evaluated using multiple corpora from different perspectives.

The contributions of this study are as follows:

- We first explore the necessity of performing cross-corpora evaluation for GEC models.
- We empirically show that the single-corpus evaluation may be unreliable.
- Our source code is published for cross-corpora evaluation so that researchers in the community can adequately and easily evaluate their models based on multiple corpora. [1]

2 Related Work

We are motivated by the issue of robustness in the parsing community. This field pre-

[1] https://github.com/tomo-wb/GEC_CCE

Proceedings of NAACL-HLT 2019, pages 1309–1314
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

viously focused on improving parsing accuracy on Penn Treebank (Marcus et al., 1993). However, robustness was largely improved by evaluation using multiple corpora including Ontonotes (Hovy et al., 2006) and Google Web Treebank (Petrov and McDonald, 2012). A situation similar to this might also occur in GEC. In other words, evaluation in GEC has relied heavily on the CoNLL-2014 benchmark, which implies that the field is overdeveloping on this dataset.

Other corpora are used for evaluation, such as KJ (Mizumoto et al., 2012) and JFLEG (Sakaguchi et al., 2017; Junczys-Dowmunt et al., 2018; Chollampatt and Ng, 2018; Ge et al., 2018; Xie et al., 2018). However, these corpora still depend on one or at most two corpora.

3 Experimental Setup

3.1 Corpora for Evaluation

Cross-corpora evaluation is discussed herein using six corpora, namely CoNLL-2014, CoNLL-2013, FCE, JFLEG, KJ, and ICNALE. The following conditions were considered when selecting corpora:

- The corpus must be used at least once in the GEC community.
- Based on the hypothesis that writers' proficiency affects the error distribution of any given text, we add a corpus with relatively low proficiency (KJ) compared to CoNLL-2014.

We explicitly describe each learner corpus as follows:

CoNLL-2014 (Ng et al., 2014), the official dataset of CoNLL-2014 shared task, is a collection of essays written by students at the National University of Singapore and is commonly used as test data for the CoNLL-2014 benchmark. This dataset contains only two essay topics.

CoNLL-2013 (Ng et al., 2013), the official dataset of CoNLL-2013 shared tasks, is commonly used as the development data for the CoNLL-2014 benchmark and contains only two essay topics.

Cambridge ESOL First Certificate in English (FCE) (Yannakoudakis et al., 2011) is a dataset containing 1,244 examination scripts of the Cambridge FCE examination. Topics and first languages (L1s) in the dataset are diversified because it contains essays for 10 topics written by non-native speakers from various countries.

JHU FLuency-Extended GUG Corpus (JFLEG) (Napoles et al., 2017) contains approximately 1,500 sentences from an English proficiency test. It contains sentences written by learners of the English language having various L1s and proficiency levels.

Konan-JIEM Learner Corpus (KJ) (Nagata et al., 2011) contains 233 essays written on 10 topics by students of a Japanese college, which are manually error-tagged and shallow-parsed.

International Corpus Network of Asian Learners of English, Written Essays (ICNALE) (Ishikawa, 2013) contains essays written by college and graduate students from ten Asian countries/regions (China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore, Taiwan, and Thailand). The original ICNALE is not error annotated. Therefore, we sampled a total number of 1,736 sentences, which are manually annotated with grammatical errors based on KJ 's annotation scheme.

Table 1 summarizes the properties of these corpora. Let N and M denote the total number of source words and sentences in a corpus, respectively. Word error rate (WER) is defined as follows:

$$\text{WER} = \frac{\sum_{m=1}^{M} d(X^m, Y^m)}{\sum_{m=1}^{M} N^m}$$

where X^m denotes each source sentence, Y^m denotes each corrected sentence, and $d(X^m, Y^m)$ denotes the edit distance between X^m and Y^m using dynamic programming.

The following conclusions are derived: (1) CoNLL-2014 has narrow coverage of topics, proficiency and L1s compared with other corporas such as JFLEG and FCE. (2) Several learner corpora are available for the evaluation of GEC models. These corpora can help investigate the performance of GEC models under different conditions.

3.2 Models

The following factors are considered while selecting our model.

- The models must be recent and commonly used.

Corpus	# sent.	# refs.	WER	# topics	Multiple L1	Multiple proficiency	Public available
CoNLL-2014	1,312	2	12.35	2	No	No	Yes
CoNLL-2013	1,381	1	14.85	2	No	No	Yes
FCE	32,199	1	12.00	10	Yes	Yes	Yes
JFLEG	747	4	20.86	Many	Yes	Yes	Yes
KJ	3,081	1	13.53	10	No	No	Yes
ICNALE	1,736	1	7.64	2	Yes	Yes	No

Table 1: Properties of evaluation corpora. Yes/No indicates whether the corpus exhibits each property in terms of multiple L1, multiple proficiency and public available.

- Each model must be implemented to have a competitive performance on CoNLL-2014.

We employed the following models based on the aforementioned factors:

LSTM: We use a bi-directional LSTM in the encoder and an LSTM with an attention mechanism in the decoder. Both the encoder and the decoder comprise two layers. The LSTM hidden state and word embedding sizes are set to be 500.

CNN: We follow the previous study (Chollampatt and Ng, 2018), namely a fully convolutional encoder–decoder architecture with seven convolutional layers. The hyperparameters used in a previous study are used herein (Chollampatt and Ng, 2018).

Transformer: Transformer is the self-attention-based model proposed by Vaswani et al. (2017). Six layers are used for both the encoder and decoder along with eight attention heads. The word embedding size is set to 1024 dimensions, and the size of position-wise feed-forward networks is set to 4096 dimensions at each inner layer.

SMT: We essentially follow the idea used in a previous study (Junczys-Dowmunt and Grundkiewicz, 2016), with some key differences. Specifically, we only use English Wikipedia for language model training and only the NUS Corpus of Learner English (NUCLE) and the Lang-8 Learner Corpora (Lang-8) for translation model training to make the experimental settings equal in all models.

3.3 Experimental Settings

We use two public datasets, namely Lang-8 (Mizumoto et al., 2011) and NUCLE (Dahlmeier et al., 2013), for training. Our pre-processing and experimental setup is similar to that reported previously (Chollampatt and Ng, 2018). In particular, a subset of NUCLE (5.4K) is utilized as the development data for selecting the model; the remaining subset (1.3M) is utilized as the training data. All the models are trained, tuned, and tested in the same way. The models are tested on each test data shown in Table 1. As an evaluation metric, we use $F_{0.5}$ score computed by applying the MaxMatch scorer (Dahlmeier and Ng, 2012) and GLEU (Napoles et al., 2015). We determine the average $F_{0.5}$ and average GLEU scores of the four models, which are trained with different random initializations, following a previously reported approach (Chollampatt and Ng, 2018).

4 Cross-Corpora Evaluation

Figure 1 shows the performance of each model sorted from best to worst based on their $F_{0.5}$ score, revealing that the performance substantially varies depending on the corpus. For example, the performance of the transformer ranges from the score of $F_{0.5}$, which is as low as 36.20 on CoNLL-2013, to as high as 60.06 on JFLEG. Notably, their rankings also considerably vary. Transformer performs best on CoNLL-2014. However, it exhibits third-best performance among FCE, KJ, and ICNALE; LSTM outperforms the other models by a large margin of up to 5.3 $F_{0.5}$ points. Some examples of the model outputs are presented in Table 2 and Table 3. Some situations are successfully corrected using transformer (Table 2), whereas it failed to perform in other situations (Table 3). The reason for difference in the model rankings cannot be generally stated because it is influenced by various factors such as the learner's proficiency, essay topic, and L1. The experimental results show, however, that discussions based on the performance on CoNLL-2014 may only hold under certain conditions.

Figure 2 shows the performance measured in GLEU having a similar trend. However, their

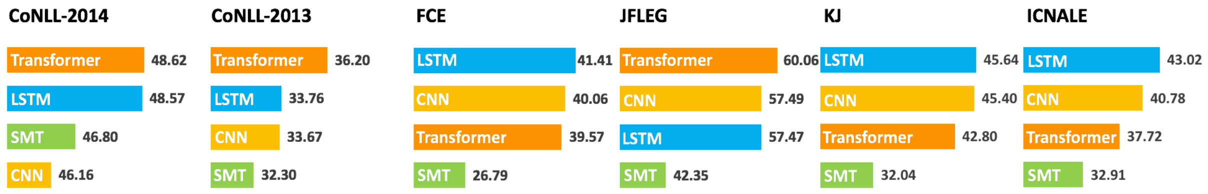

Figure 1: Average $F_{0.5}$ of the four models (trained with different random initializations), ranked best to worst.

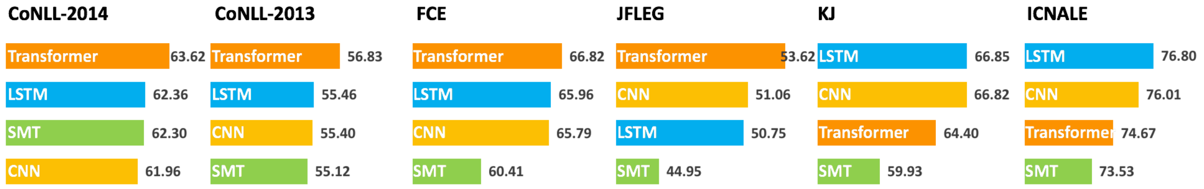

Figure 2: Average GLEU of the four models (trained with different random initializations), ranked best to worst.

rankings on FCE show different trends in Figure 1 and Figure 2. This is partly because $F_{0.5}$ and GLEU evaluate different perspectives of the models. Furthermore, evaluation data and metric must be appropriately set depending on the factors that need to be evaluated in the model.

5 Discussion

5.1 Is Diverse Single-Corpus Evaluation Sufficient?

Experimental results indicate that the benchmark single-corpus evaluation is not robust; however, more diverse corpora remain undetermined. Both JFLEG and FCE can be diverse corpora because they contain examination scripts written by language learners from all over the world. JFLEG is particularly designed to contain more diverse corpus for developing and evaluating GEC models (Napoles et al., 2017). If a diverse single-corpus evaluation suffices, the rankings of the models will remain the same. However, experimental results have shown that the model rankings on both JFLEG and FCE are different (Figure 1). Thus, single-corpus evaluation is deemed weak regardless of its diversity.

5.2 Advantage of Cross-Corpora Evaluation

This study discusses the importance of evaluating GEC models from various perspectives using multiple corpora. Multi-perspective evaluation does not necessarily mean using multiple corpora. Many aspects in a corpus can be used for analysis, such as the proficiency of the writers, essay topics, and the writer ’s native language. As a case study, we evaluate and analyze the models regarding the essay WER. Table 4 shows the performance (in precision, recall, and $F_{0.5}$) of all the models when WER is the lowest (7.64 % for ICNALE) and the highest (20.86 % for JFLEG). Transformer and LSTM outperform all the other models in the highest and the lowest error-rated corpora, respectively. Experimental results show that LSTM and transformer may be more precision-oriented and recall-oriented, respectively. Further, precision-oriented models have an advantage over recall-oriented models when a given text contains several errors, and vice versa. This knowledge enables choosing a model based on the task that has to be completed.

6 Conclusion

This study explored the necessity of performing cross-corpora evaluation for GEC models, for which the performance of several GEC models was investigated against various learner corpora. Empirical evaluation results revealed that the model performance and rankings considerably vary depending on the corpus, suggesting that a single-corpus evaluation can be unreliable. Therefore, cross-corpora evaluation should be applied to GEC models. We also published our source code for the cross-corpora evaluation framework so that researchers in the community can adequately and easily evaluate their models based on multiple corpora. Our future study will further examine the robustness of several existing evaluation metrics and explore new metrics appropriate for cross-corpora and/or cross-domain evaluation.

	Sentence
Source	Hence , some *seen* it as being considerate in keeping the genetic risk of getting the disease *in* confidential .
Reference	Hence , some **see** it as being considerate in keeping the genetic risk of getting the disease **[DEL]** confidential .
1. Transformer	Hence , some **see** it as being considerate in keeping the genetic risk of getting the disease **[DEL]** confidential .
2. LSTM	Hence , some seen it as being considerate in keeping the genetic risk of getting the disease in **confidentiality** .
3. CNN	Hence , some seen it as being considerate in keeping the genetic risk of getting the disease in **confidentiality** .

Table 2: Examples of model outputs on CoNLL-2014.

	Sentence
Source	*In* that day , the time I left school was about eleven p.m .
Reference	**On** that day , the time I left school was about eleven p.m .
1. LSTM	**On** that day , the time I left school was about eleven p.m .
2. CNN	**On** that day , the time I left school was about eleven p.m .
3. Transformer	**That** day , the time I left school was about eleven p.m .

Table 3: Examples of model outputs on KJ.

WER (%)	Low (7.64)			High (20.86)		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$
Transformer	37.69	**37.67**	37.72	67.27	**42.05**	**60.06**
LSTM	**48.68**	29.37	**43.02**	**72.97**	31.09	57.47
CNN	44.35	30.87	40.78	70.85	32.77	57.49
SMT	40.73	18.60	32.91	67.95	16.89	42.35

Table 4: Performance in precision, recall, and $F_{0.5}$ of all models on the corpora when the WER is lowest and highest.

Acknowledgments

We are grateful to the members of the Tohoku University Natural Language Processing Laboratory as well as the anonymous reviewers for their insightful comments and suggestions.

References

Shamil Chollampatt and Hwee Tou Ng. 2018. A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction. In *Proceedings of AAAI*, pages 5755–5762.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better Evaluation for Grammatical Error Correction. In *Proceedings of NAACL*, 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 *Proceedings of BEA*, pages 22–31.

Mariano Felice, Zheng Yuan, Øistein E. Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of CoNLL 2014 Shared Task*, pages 15–24.

Rachele De Felice and Stephen G. Pulman. 2008. A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English. In *Proceedings of COLING*, pages 169–176.

Tao Ge, Furu Wei, and Ming Zhou. 2018. Reaching Human-level Performance in Automatic Grammatical Error Correction: An Empirical Study. *arXiv*.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting Errors in English Article Usage by Non-Native Speakers. *Natural Language Engineering*, 12(2):115–129.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of NAACL*, pages 57–60.

Shin'ichro Ishikawa. 2013. The ICNALE and Sophisticated Contrastive Interlanguage Analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, 1:91–118.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2014. The AMU System in the CoNLL-2014 Shared Task: Grammatical Error Correction by

Data-Intensive and Feature-Rich Statistical Machine Translation. In *Proceedings of CoNLL 2014 Shared Task*, pages 25–33.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Phrase-based Machine Translation is State-of-the-Art for Automatic Grammatical Error Correction. In *Proceedings of EMNLP*, pages 1546–1556.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching Neural Grammatical Error Correction as a Low-Resource Machine Translation Task. In *Proceedings of NAACL*, pages 595–606.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2012. The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings. In *Proceedings of COLING*, pages 863–872.

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 *Proceedings of IJCNLP*, pages 147–155.

Ryo Nagata, Atsuo Kawai, Koichiro Morihiro, and Naoki Isu. 2006. A Feedback-Augmented Method for Detecting Errors in the Writing of Learners of English. In *Proceedings of COLING-ACL*, pages 241–248.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a manually error-tagged and shallow-parsed corpus. In *Proceedings of ACL*, pages 1210–1219.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground Truth for Grammatical Error Correction Metrics. In *Proceedings of ACL*, pages 588–593.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction. In *Proceedings 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 *Proceedings of CoNLL 2014 Shared Task*, pages 1–14.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of CoNLL 2013 Shared Task*, pages 1–12.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 Shared Task on Parsing the Web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language*.

Keisuke Sakaguchi, Matt Post, and Van Benjamin Durme. 2017. Grammatical Error Correction with Neural Reinforcement Learning. In *Proceedings of IJCNLP*, pages 366–372.

Ashish Vaswani, Noam Shazeer, Parmar Niki, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of NIPS*, pages 5998–6008.

Ziang Xie, Guillaume Genthial, Andrew Y. Ng, and Dan Jurafsky. 2018. Noising and Denoising Natural Language: Diverse Backtranslation for Grammar Correction. In *Proceedings of NAACL*, pages 619–628.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of ACL*, pages 180–189.