Pseudo-Error Generation for Grammatical Error Correction Based on Learner's First Language

Anonymous ACL submission

Abstract

We propose to adapt grammatical error correction (GEC) systems to the learners' first language (L1) by generating artificial errors that reflect the L1 influence. To this end, we employ two simple approaches: fine-tuning a back-translation model on L1-annotated data; and controlling the output of a back-translation model and generating artificial errors that follow the L1-dependant error type distribution. We demonstrate that, despite the simplicity of the model and the paucity of the L1-annotated data, our methods succeed in adapting GEC models to some languages. We also show that generating L1-adapted artificial errors is orthogonal to the existing method that directly adapts the GEC model to each L1. Lastly, we present an analysis of the pseudo errors generated by our models and show that they approximately capture the L1-specific error patterns.

1 Introduction

Grammatical error correction (GEC) is the task of automatically correcting grammatical errors in sentences, and various methods have been proposed to date. However, one largely overlooked aspect of GEC is that grammatical errors are highly influenced by the writer's first language (L1) (Jarvis and Odlin, 2000; Ionin et al., 2008; Yamashita and Jiang, 2010; Montrul, 2000). For instance, Ionin et al. (2008) show that Russian speakers make more errors on article use in English than do Spanish speakers, likely because Russian is an article-less language unlike English and Spanish.

The major bottleneck that prevents researchers from exploring L1-adapted GEC models is the paucity of L1-annotated data. The largest and cleanest L1-tagged learner corpus in public (First Certificate in English Corpus (FCE) (Yannakoudakis et al., 2011)) contains only a few thousand sentences for each L1, and such data do not even exist in many languages other than English due to its expensive nature. Because of this data constraint, existing studies on L1-adaptation are extremely limited and not very diverse in terms of methodology. For instance, Chollampatt et al. (2016) and Nadejde and Tetreault (2019)[1] simply fine-tune neural models on small L1-specific data, regarding L1-adaptation as a general domain adaptation problem (Daumé and Marcu, 2006; Yu et al., 2013; Luong and Manning, 2015). Rozovskaya and Roth (2010a) adapt classifier models using L1-dependent confusion sets extracted from the L1-specific data, but it is only applicable to specific types of errors (e.g. preposition errors). Therefore, much remains to be explored regarding how to make the most of the small L1-annotated data.

In this paper, instead of directly adapting GEC models, we propose to use a small number of L1-annotated sentences (e.g. 1,000 sentences) in FCE to generate a large number of L1-specific pseudo errors to perform data augmentation. To this end, we examine two approaches: the first is to fine-tune a back-translation (BT) model instead of GEC models on small L1-specific data; and the second is to control the output of the BT model and generate pseudo grammatical errors that follow the error-type distribution on L1-specific data. We demonstrate that, despite the simplicity and data paucity, these methods successfully generate pseudo errors that capture the influence of learners' first languages. Further, we show that the generated pseudo data can lead to better GEC performance overall.

2 Related Work

L1 Adaptation of GEC Models

Nadejde and Tetreault (2019) fine-tune a GEC model on learner corpora tagged with L1s or proficiency-levels and show that it improves its domain-specific performance. Similarly, Chollampatt et al. (2016) fine-tune a neural language model

[1] Notably, Nadejde and Tetreault (2019) use a superset of FCE in their experiments, i.e. Cambridge Learner Corpus (CLC) (Nicholls, 2003), which is not publicly available.

with regularization on L1-specific data and use it as a feature in their SMT-based GEC system. Rozovskaya and Roth (2010a) and Rozovskaya and Roth (2011) adapt classifier models for preposition errors by taking into account which prepositions are often confused by the specific L1 speakers. Rozovskaya et al. (2017) extend this approach and show that the L1-adaptation benefits classifier models for article and verb-agreement errors as well.

Pseudo Error Generation

Many studies employ a back-translation (BT) model (Sennrich et al., 2016b) to generate artificial errors, which "translates" grammatical sentences into ungrammatical ones. To make the BT model generate diverse errors, Xie et al. (2018) and Edunov et al. (2018) add noise to the beam search scores or sample tokens from the probability distribution and show that they are more effective than beam search decoding. Very recently, Stahlberg and Kumar (2021) have proposed a new BT model that controls its output and generates targeted types of errors. They first tag learner corpora with error types using the automatic annotation tool ERRANT (Bryant et al., 2017), and train a BT model that generates pseudo errors conditioned on those error-type tags. They show that their model generates more realistic errors and leads to better GEC performance. They have also tried adapting their model to learners' proficiency levels, and found it effective for correcting sentences written by native English speakers but not for those written by different levels of non-native speakers. Another pseudo-error generation method involves corrupting sentences with an arbitrary noise function. For instance, Rozovskaya and Roth (2010a) and Rozovskaya and Roth (2010b) generate pseudo errors by replacing correct English articles and prepositions with incorrect ones that are often misused by specific L1 speakers, which is similar in spirit to our work. However, their methods are applicable to specific types of errors only wherein the set of replacement candidates is limited. Zhao et al. (2019) and Kiyono et al. (2019) employ more random and noisy operations such as masking or shuffling words and demonstrate the effectiveness of this approach.

3 Method

We examine two simple methods for generating pseudo errors that incorporate the influence of learners' L1s. We apply both methods to a back-translation (BT) model, which converts grammatical sentences into ungrammatical ones.

Our first approach ("**Fine-tuned BT**") pre-trains a BT model on general learner corpora and fine-tunes it on small L1-specific data. In this way, we expect the model to generate pseudo errors that reflect the L1-specific error patterns, such as lexical choices and error frequency. Our second approach ("**Tagged BT**") controls the output of the BT model and generates artificial errors that follow the error type distributions on L1-specific data. This method is inspired by previous work in MT (Sennrich et al., 2016a; Johnson et al., 2017), which shows that the politeness or even the language of translations can be controlled by simply adding special tokens to the input. In this study, we control the outputs of the BT model by prepending error-type tokens such as <R:PREP> (replace preposition) to the input. We obtain these tags using the annotation tool ERRANT (Bryant et al., 2017).[2] For instance, the BT model takes "<R:SPELL> <R:PREP> I always smile at people." as an input and generates the corrupted sentence that contains the specified types of errors, e.g. "I always **simle to** people". Recently, Stahlberg and Kumar (2021) have also proposed similar yet more complex models that generate pseudo errors conditioned on the ERRANT error tags, and shown that they improve the general GEC model performance.

After training our BT models, we feed grammatical sentences into them and generate artificial errors. For Tagged BT, we sample k error tags independently according to the error type frequency distribution on L1-specific data, and prepend them to each input sentence.[3] We set k to $\lfloor \alpha N R_\ell \rfloor$, where α is a hyper-parameter, N is the number of tokens in the input sentence, and R_ℓ is the error-per-token ratio of the ℓ-specific data.

4 Experiment

4.1 Data and Experimental Setup

In our experiments, we use the learner corpora provided at BEA2019, namely Lang-8 (Mizumoto et al., 2011; Tajiri et al., 2012), FCE (Yannakoudakis et al., 2011), NUCLE (Dahlmeier et al., 2013), and W&I+LOCNESS (Granger, 1998; Yannakoudakis et al., 2018; Bryant et al., 2019).

[2] We discard OTHER tags, as their patterns are inconsistent.

[3] We also consider the part of speech tags in a sentence and avoid assigning the types of errors that are not compatible with the sentence (e.g. <R:PREP> should not be assigned to a sentence that does not contain any preposition).

Pseudo Errors	fr	ja	ru	tr
None	36.4	36.7	40.4	41.2
Standard BT	38.1	**42.0**	42.9	42.0
Tagged BT	38.2	40.4	**45.5**	**43.7**
Fine-tuned BT	**39.1**	41.4	45.2	42.1

Table 1: $F_{0.5}$ scores on FCE-ℓ-test obtained by the GEC models trained with different pseudo errors. The best scores are boldfaced.

Pseudo Errors	fr	ja	ru	tr
None	40.8	43.4	44.8	44.7
Standard BT	40.5	44.1	46.2	45.4
Tagged BT	41.3	43.5	**48.3**	**46.1**
Fine-tuned BT	**41.5**	**44.3**	47.2	45.1

Table 2: $F_{0.5}$ scores on FCE-ℓ-test obtained by the *fine-tuned* GEC models. The scores are averaged over three runs of fine-tuning the models on FCE-ℓ-dev with different random seeds. The best scores are boldfaced.

Among them, FEC is an L1-annotated learner corpus containing a few thousand sentences for each L1.[4] We split them into development and test data for each L1 $\ell \in$ {French (fr), Japanese (ja), Russian (ru), Turkish (tr)}, which we denote as FCE-ℓ-dev/test. We assign 1k sentences to FCE-ℓ-dev and use them to adapt our BT models to ℓ, by either fine-tuning a pre-trained BT model on it, or extracting the L1-specific error type distribution that Tagged BT emulates. We use the FCE data of Spanish speakers as tuning data and use them to determine the epoch size of fine-tuning the models and the hyper-parameter α, which we set to 4.0. We use the other corpora (Lang-8, W&I+LOCNESS and NUCLE) as our training data[5] for all the GEC and BT models, and the concatenation of all the sentences in FEC-{fr/ja/ru/tr}-dev as the development data. For the inputs of the BT models, we sample 1.4M sentences from Wikipedia.[6] To evaluate the effectiveness of our L1-adaptation methods, we compare the performance of GEC models that are trained with the training corpora only, or plus artificial errors generated by a standard BT model ("Standard BT"), Fine-tuned BT or Tagged BT. For all the BT and GEC models, we use the same Transformer-big architecture (Vaswani et al., 2017) to ensure fairness.[7] When we decode pseudo errors using Standard and Fine-tuned BT, we sample words from the probability distribution[8] following Edunov et al. (2018), but for Tagged BT, we use beam search and increase the number of errors by setting α to 4.0. We evaluate the GEC models using ERRANT $F_{0.5}$ scores, following recent work.

[4]Lang-8 also contains L1 information, but the data is very noisy and hence we do not use it for L1-adaptation.

[5]After pre-processing, they amount to 0.54M sentences; see Appendix A for the details.

[6]http://data.statmt.org/wmt20/translation-task/ps-km/wikipedia.en.lid_filtered.test_filtered.xz

[7]See Appendix B for the model hyper-parameters and implementation details.

[8]Otherwise, the generated errors contain very few errors and lead to much worse performance in GEC.

4.2 Result

Table 1 shows the $F_{0.5}$ scores achieved by the GEC models trained with or without different types of artificial errors. It indicates that Tagged BT improves the model performance when learners' L1 is Russian or Turkish (+2.6 and +1.7, respectively). Fine-tuned BT is also adapted well to French and Russian speakers, increasing the performance by 1.0 and 2.3. However, both models fail to improve the performance on FCE-ja-test and perform worse than Standard BT. This is presumably because the majority of the sentences in the learner corpora come from Lang-8, in which the most dominant L1 is Japanese (Brooke and Hirst, 2013).[9] Therefore, the GEC model trained on this data would be adapted to Japanese speakers already, and L1-adaptation of the pseudo errors would have little impact on the performance.

Next, we examine whether fine-tuning the GEC model itself on L1-specific data will bring further improvements, as performed by Nadejde and Tetreault (2019). Table 2 shows the results when the GEC models are pre-trained with different pseudo errors and fine-tuned on FEC-ℓ-dev.[10] It demonstrates that our models still outperform the baselines even though all the models are ℓ-adapted using the same data. This result suggests that the L1-adaptation of pseudo errors brings different benefits from adapting the GEC model itself, and both adaptation strategies are orthogonal to each other.

4.3 Analysis

To investigate whether the generated pseudo errors in fact capture the influence of L1, we compare the similarity of the error type distributions between

[9]This is because the Lang-8 corpus is scraped from the website based in Japan (https://lang-8.com).

[10]We fine-tune the GEC models with three different seeds and report the average scores. For each GEC model including the baselines, we individually tune the epoch size of fine-tuning using the tuning data, i.e. FCE-Spanish.

	L1	fr	ja	ru	tr
Standard BT	–	32.2	29.9	31.6	34.6
Tagged BT	fr	**12.9**	14.1	14.9	13.6
	ja	23.0	14.7	15.0	14.2
	ru	18.2	**11.6**	**8.4**	12.6
	tr	18.4	14.7	13.7	**10.4**
Fine-tuned BT	fr	**45.3**	43.1	44.6	**43.1**
	ja	49.2	**40.6**	42.5	43.3
	ru	50.6	41.1	**42.3**	44.4
	tr	57.0	47.9	50.7	49.7

Table 3: KL divergence of the error tag distributions between FEC-ℓ-test (column) and pseudo errors (row). For Tagged and Fine-tuned BT, the lowest values in each column are boldfaced.

	FCE-test		FCE-dev		Pseudo Errors		
	fr	ja	fr	ja	fr	ja	–
M:DET	3.7	9.4	2.6	9.8	2.1	5.7	6.3
R:VERB	7.4	5.7	7.7	4.1	2.9	2.4	2.2
R:OTHER	11.8	11.8	11.6	9.2	34.3	33.1	29.1
R:NOUN	5.1	3.6	4.7	2.9	13.0	11.4	8.7

Table 4: Error-type percentages on FCE-ℓ-test/dev and pseudo errors made by Fine-tuned and Standard BT.

the pseudo errors[11] and FEC-ℓ-test. As the similarity metric, we use the Kullback Leibler divergence (Kullback and Leibler, 1951) KL($P|Q$), where P and Q denote the error type distributions on FCE-ℓ-test and pseudo errors, respectively; and the result is shown in Table 3. First of all, it clearly shows that Tagged BT produces the most similar errors of all the models to those on the test data. It also shows that Tagged BT can adapt the errors to French, Russian and Turkish, with the adapted errors being the closest to the corresponding L1 test data.[12] Regarding Fine-tuned BT, its pseudo errors are less similar to the test data than those made by Standard BT, but are adapted to French, Japanese and Russian and capture certain important L1 influences. For instance, the percentages of the missing-determiner error found in the {fr/tr/ru/ja}-adapted pseudo errors are 2.1%, 3.7%, 5.2%, and 5.7%,[13] and these numbers clearly reflect the L1 influence; Spanish has both definite and indefinite articles like English, Turkish has an indefinite article only, and Japanese and Russian have neither of them. We also analyze other error types in Table 4, which describes the percentages of four error types on FCE-fr/ja-dev/test and pseudo errors made by Fine-tuned and Standard BT.[14] We select those errors that have the largest absolute differences between Japanese and French FCE-*dev*, and the table shows that the L1-adapted pseudo errors emulate the same magnitude relation. However, the relation of R:OTHER does not represent the L1 influence but rather the bias on FCE-dev, as its patterns are not consistent across FCE-dev and FCE-test. Hence, we expect our L1-adaptation methods to perform better as we have larger training and test data that reflect more accurate L1-specific error patterns.[15]

	Tagged BT				Fine-tuned BT			
	fr	ja	ru	tr	fr	ja	ru	tr
fr	38.2	37.8	37.5	**38.3**	**39.1**	38.6	38.4	38.4
ja	40.4	40.4	**42.3**	41.7	40.9	41.4	41.3	**41.7**
ru	44.0	42.7	**45.5**	45.0	44.1	43.5	**45.2**	42.7
tr	42.0	41.6	42.7	**43.7**	41.6	42.8	**43.1**	42.1

Table 5: $F_{0.5}$ scores on FCE-ℓ-test obtained by the GEC models adapted to ℓ. The best scores for each model are boldfaced.

Lastly, to investigate whether the L1 adaptation is indeed effective, we train GEC models with the pseudo errors adapted to different languages from L1, and Table 5 compares their performance. It shows that Tagged BT is mostly adapted to each L1 except for Japanese, for which the Russian model performs by far the best. However, as previously shown in Table 3, the Russian-adapted pseudo errors have the closest error distribution to FCE-ja-test. This indicates that generating pseudo errors that predict and mimic the error patterns on test data can be effective for improving GEC models.

5 Conclusion

We proposed a new approach to adapt GEC models to learners' L1 by generating L1-specific pseudo errors. We showed that both of the methods we explored, that is, fine-tuning a back-translation model and controlling the error types of its output, improved GEC models on L1-specific data overall. We also demonstrated that our approach brought additional benefits when combined with an existing L1-adaptation method. By analysing the generated errors, we found that our adapted BT models generated more L1-specific errors.

[11] We use the first 10k sentences to obtain the distributions.

[12] This implicitly indicates that the error tendency is consistent across FCE-ℓ-dev and -test, and that L1 influence exists.

[13] In FCE-{fr/tr/ru/ja}-test, 3.7%, 8.3%, 10.4%, and 9.4%.

[14] See Appendix C for the percentages of other error types.

[15] We also speculate that Fine-tuned BT produced much more R:OTHER errors than Standard BT as a result of the model being overfitted to the small FCE-dev data.

References

Julian Brooke and Graeme Hirst. 2013. Native language detection with 'cheap' learner corpora. In *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011)*, volume 1, pages 37–47. Presses universitaires de Louvain.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

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, Duc Tam Hoang, and Hwee Tou Ng. 2016. Adapting grammatical error correction based on the native language of writers with neural network joint models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1901–1911, Austin, Texas. Association for Computational Linguistics.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

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 the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Hal Daumé and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *J. Artif. Int. Res.*, 26(1):101–126.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

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

Tania Ionin, Maria Luisa Zubizarreta, and Salvador Bautista Maldonado. 2008. Sources of linguistic knowledge in the second language acquisition of english articles. *Lingua*, 118(4):554–576. Current emergentist and nativist perspectives on second language acquisition.

Scott Jarvis and Terence Odlin. 2000. Morphological type, spatial reference, and language transfer. *Studies in Second Language Acquisition*, 22(4):535–556.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1236–1242, Hong Kong, China. Association for Computational Linguistics.

S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Program of the 12th International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam.

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 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Silvina Montrul. 2000. Transitivity alternations in l2 acquisition: Toward a modular view of transfer. *Studies in Second Language Acquisition*, 22(2):229–273.

Maria Nadejde and Joel Tetreault. 2019. Personalizing grammatical error correction: Adaptation to proficiency level and L1. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 27–33, Hong Kong, China. Association for Computational Linguistics.

Diane Nicholls. 2003. The cambridge learner corpus: Error coding and analysis for lexicography and elt.

In *Proceedings of the Corpus Linguistics 2003 conference*, pages 572–581, Lancaster, United Kingdom.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Alla Rozovskaya and Dan Roth. 2010a. Generating confusion sets for context-sensitive error correction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 961–970, Cambridge, MA. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2010b. Training paradigms for correcting errors in grammar and usage. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 154–162, Los Angeles, California. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 924–933, Portland, Oregon, USA. Association for Computational Linguistics.

Alla Rozovskaya, Dan Roth, and Mark Sammons. 2017. Adapting to learner errors with minimal supervision. *Computational Linguistics*, 43(4):723–760.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202, Jeju Island, Korea. Association for Computational Linguistics.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. 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.

JUunko Yamashita and Nan Jiang. 2010. L1 influence on the acquisition of l2 collocations: Japanese esl users and efl learners acquiring english collocations. *TESOL Quarterly*, 44(4):647–668.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31(3):251–267.

Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide. 2013. Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7893–7897.

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 American 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.

Corpus	# Sentences
Lang-8	1,037,561
NUCLE	57,151
W&I+LOCNESS	38,691
After Preprocessing	
lang-8	497,386
NUCLE	21,539
W&I+LOCNESS	25,524
Training Data	**544,449**
FCE-fr-dev	1,000
FCE-fr-dev (non-identical)	675
FCE-fr-test	2,718
FCE-ja-dev	1,000
FCE-ja-dev (non-identical)	535
FCE-ja-test	1,361
FCE-ru-dev	1,000
FCE-ru-dev (non-identical)	623
FCE-ru-test	1,280
FCE-tr-dev	1,000
FCE-tr-dev (non-identical)	649
FCE-tr-test	1,292
FCE-dev	**4,000**
FCE-dev (non-identical)	**2,482**

Table 6: The number of sentences in each corpus

Configurations	Values
arch	Transformer (Vaswani et al., 2017)
max-tokens	8,192 (training) 1,024 (fine-tuning)
update-freq	1
seed	1,024
optimizer	Adam (Kingma and Ba, 2015)
lr	0.0005
dropout	0.3 (training) 0.1 (fine-tuning)
min-lr	1e-09
lr-scheduler	inverse_sqrt
warmup-updates	4,000
warmup-init-lr	1e-07
adam-betas	(0.9, 0.98)
max-epoch	20 (training) Tuned on FCE-Spanish (fine-tuning)
clip-norm	1.0
criterion	label_smoothed_cross_entropy (Szegedy et al., 2016)
label-smoothing	0.1
beam size	5 or 1 (sampling)

Table 7: Implementation details.

A Corpus Statistics and Preprocessing

Table 6 shows the number of sentences in each corpus we used in our experiments. We applied a few preprocessing steps to each corpus. First, we removed the sentence pairs in which more than 70% of the tokens in either the source or target sentence were composed of capital letters only (e.g. "THIS IS a PEN"). Then, we extracted the sentences written by native speakers of ℓ from the FCE corpus, and assigned 1,000 sentences to FCE-ℓ-dev and the rest to FCE-ℓ-test. Lastly, we removed the sentence pairs from the other corpora where the source and target sentences were identical, following Chollampatt and Ng (2018). When we fine-tuned BT and GEC models on FCE-ℓ-dev, we also discarded such sentence pairs, and the numbers of the remaining sentences are shown as "(non-identical)". We used the concatenation of FCE-fr/ja/ru/tr-dev (FCE-dev and FCE-dev (non-identical)) as the development data of the GEC and BT models, respectively.

B Implementation Details

Table 7 shows the model hyper-parameters and implementation settings. We used exactly the same configurations to train and fine-tune all the GEC and BT baselines and our models. We used fairseq version 0.9.0 (Ott et al., 2019) throughout our experiments.

C Error Tendency

Table 8 shows the percentages of the error types that have the 10 largest absolute differences between Japanese and French FCE-dev. It shows that Fine-tuned BT and Tagged BT emulate the same magnitude relations for the most errors. However, for some error types such as M:PUNCT, their relations are not consistent across FCE-dev (fr: 4.6%, ja: 5.9%) and FCE-test (fr: 4.5%, ja: 4.3%). This suggests that when we adapt GEC models to L1, it would be better to focus on the specific error types that are actually influenced by the native language.

D Output Examples

Table 9 shows examples of the outputs of different BT models. It indicates that Fine-tuned BTs produced noisier pseudo errors. This may be because the models were overfitted into the small L1-specific data on which they were fine-tuned. On the other hand, Tagged BTs produced more realistic errors that were specified by the error tags.

	Learner Corpora					Pseudo Errors				
	Train	FCE-dev		FCE-test		Fine-tuned		Tagged		Standard
	–	fr	ja	fr	ja	fr	ja	fr	ja	–
M:DET	6.7	2.6	9.8	3.7	9.4	2.1	5.7	4.2	11.4	6.3
R:VERB	4.3	7.7	4.1	7.4	5.7	2.9	2.4	3.6	2.1	2.2
R:OTHER	17.7	11.6	9.2	11.8	11.8	34.3	33.1	8.6	7.7	29.1
R:NOUN	2.3	4.7	2.9	5.1	3.6	13.0	11.4	4.3	3.1	8.7
R:PRON	1.3	2.9	1.2	1.8	1.2	0.3	0.2	0.9	0.4	0.3
R:VERB:FORM	2.0	3.2	1.8	2.7	2.2	0.8	0.9	2.8	2.4	1.0
R:VERB:TENSE	4.6	3.9	5.2	4.2	4.2	1.0	1.2	4.4	4.7	1.6
R:PREP	3.3	6.8	5.4	6.9	5.4	3.2	2.8	9.1	6.8	3.6
M:PUNCT	3.5	4.6	5.9	4.5	4.3	2.1	2.5	9.2	10.8	3.4
M:PREP	2.9	2.6	3.6	2.2	3.3	0.9	1.1	3.5	4.5	1.5
M:OTHER	6.0	2.3	3.2	2.5	2.4	1.7	2.0	1.9	2.1	2.5

Table 8: Error-type percentages on learner corpora and pseudo errors made by different BT models.

		Sentence
FCE	Input (Error-Corrected)	At home , we must wash our hands before lunch .
	Learner's Sentence	At home , we must wash our hands before the lunch .
Standard BT	beam search	At home , we must wash hands before lunch .
	sampling	At home , we must wash otherwise it 's lunch time .
Tagged BT	<R:PREP> <R:SPELL>	In home , we must wash our hands befor lunch .
	<M:DET>	At home , we must wash hands before lunch .
Fine-tuned BT (sampling)	French	At home , we must wash jected to eat our hands before lunch .
	Japanese	addition , we must wash hands supper .
	Russian	At home , we must wash conditiones before lunch .
	Turkish	puthome , we must Listes hands before ano our pales .
FCE	Input (Error-Corrected)	I would like to choose swimming and painting .
	Learner's Sentence	I would willingly like to choose swimming and painting .
Standard BT	beam search	I would like to choose swimming and painting .
	sampling	I would architecture . . and painting . ally .
Tagged BT	<R:VERB>	I would like to choice swimming and painting .
	<U:NOUN>	I would like to choose swimming color and painting .
Fine-tuned BT (sampling)	French	I would concera likes swimming and guy ided painting .
	Japanese	I would clever , order bou to processes ' like to choose swimming and painting .
	Russian	I would spelling to flowers swimming and painting warf .
	Turkish	I would exhiat Tide swimming , paint .

Table 9: Examples of pseudo errors