Beyond Grammatical Error Correction: Improving L1-influenced research writing in English using pre-trained encoder-decoder models

Gustavo Zomer, Ana-Frankenberg Garcia
Centre for Translation Studies, University of Surrey
{g.zomer, a.frankenberg-garcia}@surrey.ac.uk

Abstract

In this paper, we present a new method for training a writing improvement model adapted to the writer's first language (L1) that goes beyond grammatical error correction (GEC). Without using annotated training data, we rely solely on pre-trained language models fine-tuned with parallel corpora of reference translation aligned with machine translation. We evaluate our model with corpora of academic papers written in English by L1 Portuguese and L1 Spanish scholars and a reference corpus of expert academic English. We show that our model is able to address specific L1-influenced writing and more complex linguistic phenomena than existing methods, outperforming what a state-of-the-art GEC system can achieve in this regard. Our code and data are open to other researchers[1].

1 Introduction

Writing for international readerships can be challenging for scholars whose working language is not English (Flowerdew, 2019), even when they are reasonably proficient in the language. However, existing research on tools for improving writing focuses mostly on Grammatical Error Corrrection (GEC), and often relies on college-level corpora as a benchmark for lower-English proficiency levels. While GEC can address more straightforward grammar mistakes, it does not adequately address fluency and more complex linguistic issues (Napoles et al., 2017). Moreover, few studies address the effects of first-language (L1) transfer on writing (Nadejde and Tetreault, 2019).

Another limitation is that state-of-the-art methods in GEC such as neural machine-translation-based (NMT) approaches (Sennrich et al., 2016a) and transformer-based sequence-to-sequence models (Vaswani et al., 2017) typically require a large amount of pseudo-errors generated from monolingual data using rule-base corruption (Zhao et al., 2019), back-translation (Kiyono et al., 2019), or round-trip translation (Lichtarge et al., 2019). These methods tend to introduce errors with limited diversity, most of which involve spelling and grammar.

There have been recent attempts to eliminate the time-consuming pre-training step by employing pre-trained transformer models. Alikaniotis and Raheja (2019) used pre-trained transformer models in a language-model setting, and Katsumata and Komachi (2020) fine-tuned BART (Lewis et al., 2020) with a small corpus of annotated sentences. Both approaches achieved results comparable to models trained with millions of sentences.

This study proposes a method for improving L1-influenced English texts beyond GEC through the use of pre-trained encoder-decoder models. Our approach uses parallel corpora of English aligned with English that has been machine-translated from Portuguese and Spanish as training data that emulates L1-influenced writing. The models generate sentences that have a higher level of acceptability and are more linguistically diverse than what a state-of-the-art GEC system can achieve. We also propose new metrics for evaluating improvement beyond correctness.

2 Related work

Chollampatt et al. (2016) adapted a neural language model for three different L1s and Nadejde and Tetreault (2019) adapted a general-purpose neural GEC system to the writer's proficiency level and L1. Both studies were conducted on student corpora. Takahashi et al. (2020) proposed a method to generate more realistic pseudo errors by considering learners' tendency for errors, but only adjusted the probability of grammatical and spelling mistakes.

More recently, Zhou et al. (2020) proposed a novel data-synthesis method to generate error-corrected sentence pairs based on a pair of machine-

---

[1] https://github.com/gzomer/BeyondGEC

2534

translation models of different qualities. However, good quality NMT translation may not be good enough as a reference.

3 Method

Our method consists of fine-tuning a pre-trained encoder-decoder transformer model with parallel corpora of English and machine-translated English.

3.1 Parallel corpora compilation

We create the parallel corpora by machine-translating into English the non-English sentences of a parallel corpus containing English. Our assumption was that machine-translated text can be used as a proxy for L1-influenced English, as it tends to render more literal translations, preserving many lexical and syntactical features of the source language. This can mimic the output of writers using English as a second language (L2) who transfer words and grammar from their L1 when writing in English. The example below illustrates this. Sentence (a), from the Brazilian Academic Corpus of English (BrACE) (Pinto et al., 2021), exemplifies the common problem among Brazilians writing in English of using *realization* to mean *undertaking*. This is likely to be due to the direct translation of the Portuguese false cognate 'realização', as shown in sentence (b), a Portuguese sentence from the EurLex corpus (Baisa et al., 2016). Sentence (c) shows how the same problem occurs when (b) is submitted to MT, but not in the reference translation in (d).

a. BrACE: For the *realization* of the initial stage of the project, nationwide semi-presential extension courses were proposed [...]

b. PT Source Text: O próprio presidente do governo regional afirmou que estão garantidos os fundos de Bruxelas para a *realização* das referidas obras.

c. PT-EN MT: The president himself of the regional government said that Brussels funds are guaranteed for the *realization* of these works.

d. PT-EN Reference Human Translation: The President of the Regional Government himself has said that funds from Brussels are guaranteed to enable the projects to be started.

This study used Spanish and Portuguese as L1. We selected the EUR-Lex Portuguese, Spanish and English subcorpora as a starting point. We corrected alignment issues using the length-based alignment algorithm from Gale and Church (1993).

Next, we cleaned the corpora by removing duplicated sentences and sentences in other languages using the `langid`[2] package. We also removed sentences with fewer than 10 words to filter out titles. Empty alignment units were discarded.

Finally, we machine translated the Spanish and Portuguese corpora into English using the Open-MT `roa-en` model[3] (Tiedemann and Thottingal, 2020). We obtained 260k sentences[4] for each: (a) the EN$_{PT}$-EN corpus of Portuguese-machine-translated-into-English aligned with English and (b) the EN$_{ES}$-EN corpus of Spanish-machine-translated-into-English aligned with English.

3.2 Model

We used the Text-to-Text Transfer Transformer (T5) model (Raffel et al., 2020) as our encoder-decoder. In T5, each task is treated as a language-generation task where the model is conditioned to generate the correct output based on a textual prompt included in the input sequence (Raffel et al., 2020). We fine-tuned T5 by using the parallel corpora we created, with the machine-translated sentence as the input and the English reference as the output. We conditioned the input sentence on the writers' L1 by prepending the task definition improve_english <L1>: for each input, where <L1> is replaced by the writer's L1.

3.3 Pre- and post-processing

We added a pre-encoding and post-decoding step in our model to preserve out-of-vocabulary (OOV) tokens, as academic and scientific texts include a diversity of mathematical and other non-alphanumeric symbols. Adding symbols to the tokenizer would significantly increase the model vocabulary and decrease its performance. OOV tokens were replaced with special tokens appended with an id **[KEEP_ID]**, and then restored with the original token post-decoding.

4 Evaluation

We evaluated our model using four different metrics: grammaticality, acceptability, and lexical and syntactical diversity.

---

[2]https://github.com/saffsd/langid.py
[3]Available in the Hugging Face library (Wolf et al., 2020)
[4]Corpora available on https://github.com/gzomer/BeyondGEC

4.1 Grammaticality

We define our grammaticality metric as the accuracy of a RoBERTa[5] (Liu et al., 2019) classifier trained on the CoLA corpus (Warstadt et al., 2019), which contains sentences paired with grammatical acceptability judgments, following a similar approach to that of Krishna et al. (2020).

4.2 Acceptability

We measure acceptability based on the assumption that probability is related to naturalness, as shown in Lau et al. (2017). We used SLOR (Pauls and Klein, 2012), as it is particularly effective in neutralizing both sentence length and word frequency, and has yielded the best results in a comparative study of different metrics (Lau et al., 2017). SLOR is calculated as the normalized difference between the sentence log-probability and the uni-gram sentence probability, as shown below:

$$SLOR = \frac{\log P(s) - \log P_u(s)}{|s|} \quad (1)$$

To calculate the probabilities, we built the Expert Academic Corpus of English (ExpACE) corpus, which consists of 10 million words from over 1200 highly cited papers published in high-impact journals (based on h5 index) in eight different subject domains. We removed references, tables, and replaced citations with **(CIT)**. We fine-tuned GPT-2 with ExpACE on a Tesla T4 GPU for 3 epochs with

---

[5] We fine-tuned Hugging Face `RoBERTa-base` on a Tesla T4 GPU for 3 epochs with batch size of 32 and using Adam optimizer with a learning rate of $2e^{-5}$

batch size of 8 and using Adam optimizer with a learning rate of $3e^{-5}$ to calculate sentence probabilities, and built a subword uni-gram model using byte-pair encoding (BPE) (Sennrich et al., 2016b) as the tokenizer. We multiplied each BPE token probability to approximate the word probability.

4.3 Lexical and syntactic diversity

As discussed in the introduction, most GEC systems focus on correctness and minor-edits (Napoles et al., 2017). However, L1-influenced texts may underuse and overuse certain vocabulary (Pinto et al., 2021), and may require more complex syntactical transformations to read more fluently. Therefore, we propose two new measures for evaluating linguistic diversity at the lexical and syntactical level.

The lexical diversity metric is measured by the sum of lexical changes with different lemmas (e.g. changing *provoked* to *led*) normalized by the sentence length. Models focusing mostly on grammaticality (e.g. fixing verb tense from *show* to *shown*) will score lower on this metric. We calculated lexical diversity by first word-aligning the original and improved sentence using SimAlign (Jalili Sabet et al., 2020). Lemmas were extracted using WordNet from the `nltk` package.

We also introduce a syntactical diversity metric, defined as the normalized difference between the word alignments of the original and improved sentence. The idea behind this metric is to capture reordering of words and phrases in a sentence, as non-conventional order can decrease readability (Wallwork, 2016).

Model	Sentence	SxD	LxD
Source	A company that has a good planning, with well-defined action plans, indicators and responsible persons, will have a favorable condition.		
GECToR	A company that has good planning, with well-defined action plans, indicators and responsible persons, will have a favorable condition.	0.000	0.000
Our model	A well-planned company with well-defined action plans, indicators and responsible persons will have a favorable condition.	5.000	0.000
Source	Another stream of thought about authenticity is initiated by the sociologist MacCannell (1973) in the early 1970s.		
GECToR	Another stream of thought about authenticity was initiated by the sociologist MacCannell (1973) in the early 1970s.	0.000	0.000
Our model	The sociologist MacCannell (1973) initiated a further stream of thought about authenticity in the early 1970s.	5.737	0.229
Source	Thus, in order to ensure the uses in a sustainable manner, it is important to manage water resources through structural and non-structural measures.		
GECToR	Thus, in order to ensure the use in a sustainable manner, it is important to manage water resources through structural and non-structural measures.	0.000	0.196
Our model	In order to ensure sustainable use, it is therefore important to manage water resources through structural and non-structural measures.	4.667	0.218

Table 1: Examples comparing the output of our model with GECToR

5 Experiments

In this section, we describe the two corpora compiled for evaluating our model and the experimental setup used in the evaluation.

5.1 Evaluation corpora

The first corpus is a larger version of the 1-million word Brazilian Academic Corpus of English (BrACE) (Pinto et al., 2021), consisting of 14 million words of journal articles published in English in Brazilian journals. The second corpus is the Latin-American Academic Corpus of English (LACE) corpus, containing 13 million words of research articles published in English in journals from Spanish Latin America. Both corpora were compiled using a balanced sample of seven broad subject areas downloaded from Scientific Electronic Library Online (SciELO). We cleaned both corpora by removing headers, references, footnotes, figures and tables. We also replaced citations with **(CIT)**.

5.2 Experimental setup

We experimented with a small T5 model version with 60M trainable parameters and a larger version with 220M parameters[6]. We fine-tuned each model for 3 epochs using Adam optimizer with a learning rate of $5e^{-5}$ and a batch size of 16. We compared our results with GECToR (Omelianchuk et al., 2020), a current state-of-the GEC model.

During inference, we performed beam search with a beam size of five. We also added a τ threshold above which we accept the candidate sentence by comparing the original sentence and the candidate sentence probability evaluated using the GPT-2 language model fine-tuned with ExpACE (see Section 4.2), similar to the approach used in (Krishna et al., 2020). As this step is applied post-inference, we applied τ filtering to GECToR in order for the results to be comparable. In our experiments, we empirically found a τ of 0.05 as a good balance between precision and recall.

6 Results and Discussion

The results for BrACE are presented in Table 2 and for LACE in Table 3, where we present the average change in grammaticality (GRM), average change in acceptability (ACP), lexical diversity (LxD), and syntactical diversity (SxD) on a sample of 20k sentences of each corpus.

---

[6] Hugging Face `t5-small` and `t5-base`

Model	GRM	ACP	LxD	SxD
GECToR	0.016	-0.009	0.048	0.145
GECToR + τ	**0.067**	0.297	0.162	0.347
L1-aware-small	0.036	0.447	0.226	0.815
L1-aware-large	0.046	**0.517**	**0.265**	**1.089**

Table 2: Results for BrACE corpus

Model	GRM	ACP	LxD	SxD
GECToR	0.016	-0.021	0.051	0.172
GECToR + τ	**0.069**	0.321	0.168	0.336
L1-aware-small	0.039	0.428	0.219	0.841
L1-aware-large	0.048	**0.544**	**0.263**	**1.158**

Table 3: Results for LACE corpus

Our results show that both small and large models generate sentences that have a greater degree of acceptability and are more lexically and syntactically diverse by a large margin in comparison with the GECToR baseline. On the other hand, GECToR performed better on grammatically. This was not unexpected, as our model was not optimized for grammatically and was trained with fewer sentences.

Table 1 exemplifies the differences between minimal-edits from GECToR and the more substantial changes from our model. As shown in the first sentence, our model is able to improve *that has a good planning* to *well-planned*, whereas GECToR only deleted an article.

In addition, we analysed the ability of the models to capture L1-influence. We extracted the 1000 most frequent 2-,3-,4-grams of the CoPEP corpus of journal articles written in Portuguese (Kuhn, 2017), translated them into English and compared their frequency in BrACE and ExPACE, to test whether the CoPEP n-grams were overused in BrACE.

Type	% of changes with errors
Over-correction	6.69%
Extra/Missing symbols	5.28%
Semantical	3.17%
Grammatical	2.82%
Possessive	1.41%
Terminology	1.41%
Word choice	1.41%
Word order	1.06%
Comma	1.06%
Spelling	1.06%
Repetitive	0.70%

Table 4: Error analysis

We found that 74% of the n-grams were indeed overused in BrACE when compared with ExPACE. Our model revised 23% of these, while GECToR only tackled 14%. In doing so, overall n-gram overuse was reduced by 34% with our model against 19% with GECToR.

A preliminary error analysis of the output of our model was conducted using a sample of 100 sentences from BrACE that were submitted to our system. We annotated each sentence pair with ER-RANT (Bryant et al., 2017) and manually classified 284 changes in one or more categories of errors. Table 4 shows that the most common type of issue is overcorrection, especially occurring among cohesive devices (for example, changing *furthermore* to *moreover*). Removing symbols or adding unnecessary ones was the second most common type of problem (e.g., removing punctuation). Although changes in terminology were uncommon, further research should explore approaches for keeping them unchanged.

Although these issues could be mitigated by increasing τ filtering, it would also filter out relevant suggestions. Instead, we suggest further research on incorporating local biases on specific tokens, giving either a boost or reduction on the likelihood of changing tokens. The GECToR approach of tagging sentences instead of rewriting them allows for the use of bias to keep tokens unchanged. However, this restricts the flexibility of more complex and linguistic diverse changes, as seen in our results. We argue that future research should incorporate the benefits of both approaches in a single method, achieving high-flexibility with controlled results.

7 Final Remarks

We introduced a new method for improving L1-influenced academic English by using a pre-trained encoder-decoder transformer. We showed that our model generates sentences with a higher acceptability and that are more linguistically diverse than a state-of-the-art GEC model. However, further research is needed to assess the extent to which our model over-corrects or introduces errors.

The approach taken in this study can be extended to other L1s by modifying existing parallel corpora with machine translation. It can also be extended to other domains, beyond academic, as our method does not rely on annotated corpora. It is evident that more research is needed beyond grammatical correction, including the development of corpora and models addressing more linguistically diverse issues.

References

Dimitris Alikaniotis and Vipul Raheja. 2019. The unreasonable effectiveness of transformer language models in grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 127–133, Florence, Italy. Association for Computational Linguistics.

Vít Baisa, Jan Michelfeit, Marek Medveď, and Miloš Jakubíček. 2016. European Union Language Resources in Sketch Engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2799–2803, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

John Flowerdew. 2019. The linguistic disadvantage of scholars who write in english as an additional language: Myth or reality. *Language Teaching*, 52(2):249–260.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Satoru Katsumata and Mamoru Komachi. 2020. Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 827–832, Suzhou, China. Association for Computational Linguistics.

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.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Tanara Zingano Kuhn. 2017. A design proposal of an online corpus-driven dictionary of portuguese for university students.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. 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 3291–3301, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

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.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Paula Tavares Pinto, Geraint Paul Rees, and Ana Frankenberg-Garcia. 2021. Identifying collocation issues in English L2 research article writing. In Maggie Charles and Ana Frankenberg-Garcia, editors, *Corpora in ESP/EAP Writing Instruction: Preparation, Exploitation, Analysis*. Routledge.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Yujin Takahashi, Satoru Katsumata, and Mamoru Komachi. 2020. Grammatical error correction using pseudo learner corpus considering learner's error tendency. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 27–32, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenc of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, page 6000–6010.

Adrian Wallwork. 2016. *English for writing research papers*. Springer, London, England.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. 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 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.

Wangchunshu Zhou, Tao Ge, Chang Mu, Ke Xu, Furu Wei, and Ming Zhou. 2020. Improving grammatical error correction with machine translation pairs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 318–328, Online. Association for Computational Linguistics.