Grammatical Error Correction: Are We There Yet?

Muhammad Reza Qorib and **Hwee Tou Ng**
Department of Computer Science, National University of Singapore
{mrqorib, nght}@comp.nus.edu.sg

Abstract

There has been much recent progress in natural language processing, and grammatical error correction (GEC) is no exception. We found that state-of-the-art GEC systems (T5 and GECToR) outperform humans by a wide margin on the CoNLL-2014 test set, a benchmark GEC test corpus, as measured by the standard $F_{0.5}$ evaluation metric. However, a careful examination of their outputs reveals that there are still classes of errors that they fail to correct. This suggests that creating new test data that more accurately measure the true performance of GEC systems constitutes important future work.

1 Introduction

Grammatical Error Correction (GEC) is a task that has many real-world applications, such as proofreading, assisting language learners (Knutsson et al., 2003; Chollampatt et al., 2016; Nadejde and Tetreault, 2019), assisting children with developmental language disorder (Balthazar et al., 2020), etc. In order to measure the accuracy of a GEC system, multiple evaluation metrics have been proposed in the past. Ever since the CoNLL-2014 shared task (Ng et al., 2014), the $F_{0.5}$ metric has been used as the standard evaluation metric for GEC. The $F_{0.5}$ score has been found to have a better correlation with human judgment compared to F_1 and other metrics (Grundkiewicz et al., 2015; Napoles et al., 2015; Chollampatt and Ng, 2018).

As $F_{0.5}$ is a reference-based metric, its computation relies on comparing a system's output sentences to human's corrected sentences (called references henceforth). Typically, there are multiple ways to correct an input sentence. For example, the CoNLL-2014 (Ng et al., 2014) official test set contains 2 references for each input sentence, and the BEA-2019 (Bryant et al., 2019) test set contains 5 references. Choshen and Abend (2018) even predict that a short sentence may have more than 1,000 valid corrections. Thus, this evaluation scheme tends to underestimate the performance of a GEC system. Rozovskaya and Roth (2021) have reported that GEC systems can obtain higher scores if evaluated using references that are closer to a system's outputs.

The limited references mean that humans may also not reach 100% $F_{0.5}$ performance, because their corrections need not be the same as the references. Bryant and Ng (2015) was the first to attempt to measure human performance on the CoNLL-2014 test set by adding 8 new references for each sentence, so that each sentence has 10 references. Then, they calculate each annotator's performance by comparing his corrections to the other 9 annotators. The average $F_{0.5}$ score of 72.58% from all annotators is then considered as the estimated human-level performance. To compare a GEC system's performance against human performance, a similar procedure is applied by taking the average of the system's performance on the 10 sets of 9-annotator references.

Recently, much progress has been made in natural language processing. For example, for some question answering datasets (such as SQuAD (Rajpurkar et al., 2016) and CoQA (Reddy et al., 2019)) and natural language understanding test suites (SuperGLUE (Wang et al., 2019)), superhuman performance has been reported where the performance of the best NLP system exceeds human performance. A natural question to ask is: how does the performance of the latest state-of-the-art (SOTA) GEC systems compare to human performance? In this paper, we first attempt to answer this question. We found that state-of-the-art GEC systems (T5 (Rothe et al., 2021) and GECToR (Omelianchuk et al., 2020)) outperform humans by a wide margin on the CoNLL-2014 test set, as measured by the $F_{0.5}$ metric. However, a careful examination of their outputs reveals that there are still classes of errors that they fail to cor-

rect. This suggests that creating new test data that more accurately measure the true performance of GEC systems constitutes important future work.

2 Are We There Yet?

At the time of writing this paper, the best published GEC system using a sequence-to-sequence approach is T5-XXL, and the best published GEC system using a sequence-tagging approach is the ensemble of GECToR XLNet and GECToR RoBERTa. In this paper, we use both systems, but with T5-XXL replaced by T5-Large due to resource constraints, since T5-Large has 770M parameters while T5-XXL has 11B parameters. Our evaluation is based on the scheme proposed by (Bryant and Ng, 2015), using 10 sets of human annotations of the CoNLL-2014 official test set. The $F_{0.5}$ scores of T5 and GECToR are shown in Table 1. We find that both systems outperform humans by a wide margin – T5 outperforms humans by 6.05 points and GECToR outperforms humans by 8.33 points.

Model	$F_{0.5}$ Score
Human	72.58
T5	78.63
GECToR	**80.91**

Table 1: The performance of SOTA GEC systems in comparison to human performance.

However, even though the scores in Table 1 show that SOTA GEC systems have outperformed humans, when we check the distribution of per-sentence $F_{0.5}$ scores, we find that both T5 and GECToR generate sentences with a zero $F_{0.5}$ score even comparing to 9 references. Specifically, on the CoNLL-2014 test set, T5 (GECToR) completely fails to correct 9.1% (12.8%) of the sentences (Figure 1). On the BEA-2019 development set, the proportion is higher, with 27% of the sentences resulting in a zero $F_{0.5}$ score for T5, and 29.9% for GECToR (Figure 2)[1].

3 Current Weaknesses

To understand what causes the low scores of the top-performing GEC systems, we examine a sample of 100 sentences from the systems' outputs on the BEA-2019 development set, starting from sentences with the lowest $F_{0.5}$ score. We found that even though T5 and GECToR generally produced good corrections, they also sometimes made obvious mistakes that humans will not make.

3.1 Unnatural Phrases

Source	The first place was gotting by us .
T5	The first place was **got** by us .
GECToR	The first place was gotting by us .
Target	**We won first place** .

Table 2: An example of GEC systems failing to fix unnatural phrases.

In the example in Table 2, GECToR completely fails to produce any corrections, with the misspelling uncorrected. T5 successfully makes the sentence grammatical by applying the appropriate edit, but it fails to make it sound natural. Similarly, both systems also fail to correct the sentence in Table 3. One possible reason is that this kind of sentences where the target sentence corrects unnatural phrases happen rarely in the training data, since the human annotators are expected to make minimal edits to make the sentence grammatically correct, instead of making the sentence sounds more natural, like the goal of the JFLEG dataset (Napoles et al., 2017).

Source	I like personality with childlike , so I like children .
T5	I like personality with childlike , so I like children .
GECToR	I like personality with childlike , so I like children .
Target	I like **childlike people**, so I like children .

Table 3: Another example of GEC systems failing to fix unnatural phrases.

3.2 Sentence Structure

In the example in Table 4, the systems fail to detect that the object of the sentence consists of multiple items. GEC systems sometimes fail to detect the subject-verb relationship when the subject or object is in a form of a complex phrase instead of a single word. This observation is in line with Mita and Yanaka (2021) who reported that the standard Transformer-based GEC model has difficulties in fixing subject-verb agreement on error patterns that do not appear in the training data, even in simple settings with limited vocabulary and syntax.

[1]However, note that since the BEA-2019 development set has only one reference per sentence, there is a greater chance of underestimating system performance.

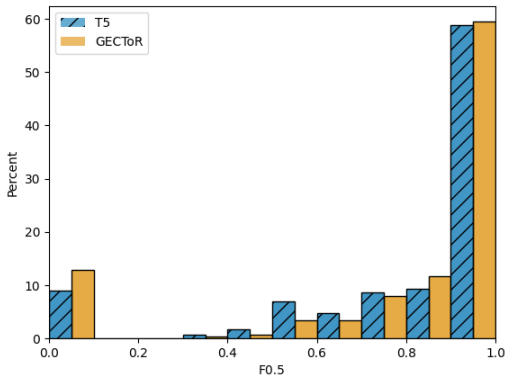

Figure 1: Distribution of per-sentence $F_{0.5}$ scores on the CoNLL-2014 test set.

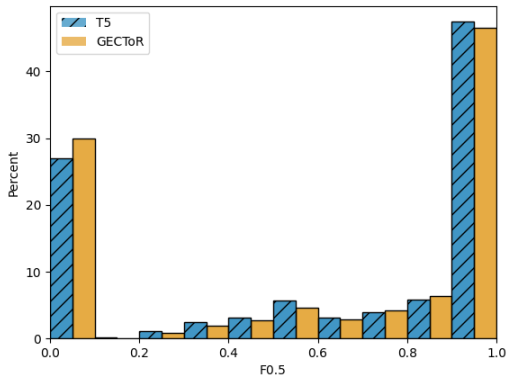

Figure 2: Distribution of per-sentence $F_{0.5}$ scores on the BEA-2019 development set.

Source	There are a little kitchen , a great bedroom , a bathroom with shower but without bath and a cool living - room .
T5	There **is** a little kitchen , a great bedroom , a bathroom with shower but without bath and a cool living - room .
GECToR	There **is** a little kitchen , a great bedroom , a bathroom with shower but without bath and a cool living - room .
Target	There are a little kitchen , a great bedroom , a bathroom with shower but without bath and a cool living - room .

Table 4: An example of GEC systems failing to detect the sentence structure.

Source	The implications for the beef industry could be rather serious , where everybody to boycott beef products .
T5	The implications for the beef industry could be rather serious , where everybody **boycotts** beef products .
GECToR	The implications for the beef industry could be rather serious , where everybody to boycott beef products .
Target	The implications for the beef industry could be rather serious , **were** everybody to boycott beef products .

Table 5: An example of GEC systems failing to understand a sentence's meaning.

3.3 Sentence Comprehension

The example in Table 5 requires a full comprehension of the sentence to make the right correction. Another example in Table 6 also shows that the GEC systems fail to understand that getting along with his or her companion is what the writer wants to convey, given the context of wanting to enjoy a trip.

3.4 Error Rates

On the BEA-2019 development set, we analyze the effect of error rate (the percentage of erroneous tokens in a source sentence) on GEC systems' per-sentence $F_{0.5}$ score. The error rate is computed as the number of tokens in a source sentence that are to be deleted or substituted by the edits in the gold reference, divided by the total number of tokens in the sentence. Next, we remove the outliers, which are sentences with error rates $\pm$ 3 standard deviations from the mean, and we end up removing 1.64% of the sentences. The relationship between the per-sentence $F_{0.5}$ score of a sentence and its error rate (rounded to the nearest 0.05) is presented in Figure 3. We observe that GEC systems perform better on sentences with low error rates (less than 0.05).

3.5 Long Sentences

We also analyze the effect of sentence length on a GEC model's performance. On the BEA-2019 development set, we count the number of words in each source sentence and obtain the per-sentence $F_{0.5}$ score of the sentence. Then, we group the sentences based on sentence length in step size of 5 words (0–4 words, 5–9 words, etc). We apply the same outlier removal procedure in Section 3.4, eliminating 1.25% of the sentences.

We observe that top GEC systems have difficulties in generating accurate corrections for long sen-

Source	In order to enjoy a trip to Mexico I suggest that the traveler find a manner to get alone with his or her companion.
T5	In order to enjoy a trip to Mexico **,** I suggest that the traveler find a **way** to get alone with his or her companion.
GECToR	In order to enjoy a trip to Mexico **,** I suggest that the traveler find a manner to get alone with his or her companion.
Target	In order to enjoy a trip to Mexico **,** I suggest that the traveler **finds** a **way** to get **along** with his or her companion.

Table 6: Another example of GEC systems failing to understand a sentence's meaning.

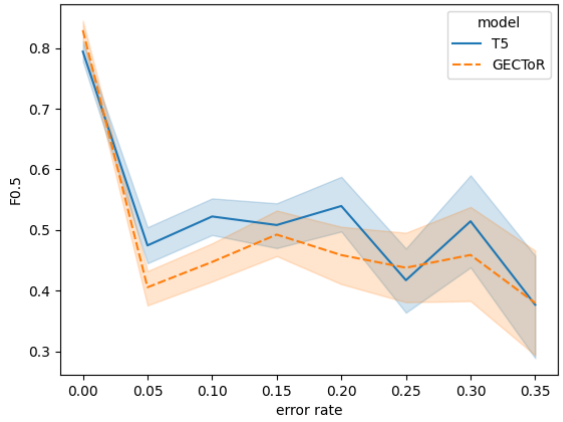

Figure 3: Per-sentence $F_{0.5}$ scores of top GEC models for different error rates. The straight line segments connect the mean values and the shaded region denotes the variance.

tences. As shown in Figure 4, the per-sentence $F_{0.5}$ shows a downward trend with increasing sentence length.

3.6 Cross-Sentence Context

Most of the current GEC systems are working at the sentence level. Cross-sentence GEC has not been given enough attention, and only a few papers have pursued this direction (Chollampatt et al., 2019; Yuan and Bryant, 2021). However, some sentences indisputably require cross-sentence context to correctly fix them. For example, the correction for the sentence in Table 7 requires knowing the context to realize that there is a misspelling in the sentence. Typical errors that require cross-sentence context to correct include pronoun agreement and tense agreement (Table 8). However, the errors can

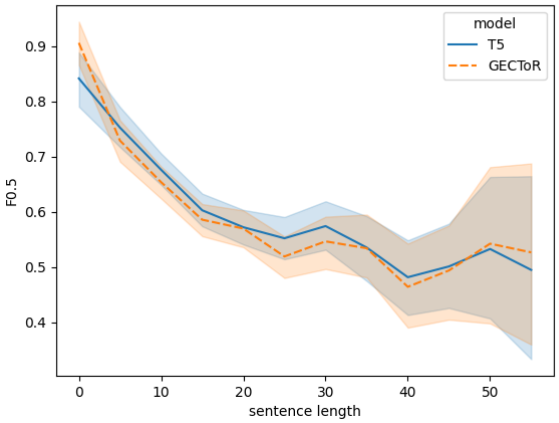

Figure 4: Per-sentence $F_{0.5}$ scores of top GEC models for different sentence lengths. The straight line segments connect the mean values and the shaded region denotes the variance.

be of many kinds, in the form of sentence structure, word choice, misspelling, and many more.

3.7 Adapting to a Different Domain

The common GEC training corpora, such as NUCLE (Dahlmeier et al., 2013), FCE (Yannakoudakis et al., 2011), Lang-8 (Mizumoto et al., 2011), and W&I+LOCNESS (Granger, 1998; Yannakoudakis et al., 2018), and GEC test corpora, such as CoNLL-2014, BEA-2019, and JFLEG, originated from essays written by English as a second language (ESL) authors. Those datasets are more situated in an academic setting. To know how GEC models perform in a different domain, we evaluate the models on the CWEB datasets (Flachs et al., 2020) without additional training (Table 9)

The CWEB dataset is made from website data to represent language correction "in the wild". The CWEB dataset consists of two subsets: CWEB-S which has source sentences from government, educational institution, and museum websites, and CWEB-G from general websites. When tested on CWEB, the $F_{0.5}$ scores of both T5 and GECToR drop drastically from their CoNLL-2014 test scores, especially for T5. These reductions indicate that current top GEC models are not robust to domain change, even though GEC models are expected to be able to correct all kinds of sentences.

4 Moving Forward

As we have seen examples of how SOTA GEC models fail to correct sentences that are easy for hu-

Context	Is there a future for Privately owned cars ? To be honest , I am not sure . Although privately oned cars are more and more popular , ...
Source	I would say that most probably private care are not " sustainables " in the long term , ...
T5	I would say that most probably private care **is** not " **sustainable** " in the long term , ...
GECToR	I would say that **,** most probably **,** private care **is** not " **sustainable** " in the long term , ...
Target	I would say that **,** most probably **,** private **cars** are not " **sustainable** " in the long term , ...

Table 7: An example of a sentence that requires cross-sentence context to correct.

Context	For example there are girl she is in my class she is beautiful . I love her look . Her eyes looks as the sun .
Source	But she is ignored me .
T5	But she ~~is~~ ignored me .
GECToR	But she ~~is~~ ignored me .
Target	But she ~~is~~ **ignores** me ..

Table 8: Another example of a sentence that requires cross-sentence context to correct.

mans, we can conclude that GEC models have not actually outperformed humans in practice. However, the current top GEC systems already reach high scores on the standard benchmarks. Thus, we argue that it is important to create a new test set that contains more sentences that pose challenges to SOTA GEC models but can be easily corrected by humans. The test set may emphasize on sentences with complex noun phrase subject/object, sentences that require full comprehension to be corrected, sentences that contain many grammatical errors, sentences that are long, sentences that require cross-sentence context, etc. With a harder test set, we can then more clearly assess how far we are from considering GEC as a solved task.

5 Conclusion

In this paper, we have reported that the top GEC systems using the sequence-to-sequence approach (T5) and sequence-tagging approach (GECToR) have produced $F_{0.5}$ scores that exceed that of humans, as measured by Bryant and Ng (2015). Based on qualitative analysis of their outputs, we conclude that even though GECToR and T5 achieve $F_{0.5}$ scores higher than that of human, they have not outperformed humans in practice as they still fail to correct sentences that can be easily corrected by humans. We also report our qualitative analysis on the weaknesses of current top GEC models, which point to directions for future research. Lastly, we argue that a new test set that emphasizes harder sentences is needed to evaluate the progress of GEC as a field.

Test data	T5	GECToR
CoNLL-2014	65.07	66.05
CWEB-G Dev	38.91	40.79
CWEB-G Test	39.80	42.67
CWEB-S Dev	27.37	37.63
CWEB-S Test	28.51	33.07

Table 9: The performance of T5 and GECToR on the CoNLL-2014 test set and CWEB.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-RP-2019-014). The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

Catherine H Balthazar, Susan Ebbels, and Rob Zwitserlood. 2020. Explicit grammatical intervention for developmental language disorder: Three approaches. *Language, Speech, and Hearing Services in Schools*, 51:226–246.

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 and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707, Beijing, China. 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 reassessment of reference-based grammatical error correction metrics. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2730–2741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shamil Chollampatt, Weiqi Wang, and Hwee Tou Ng. 2019. Cross-sentence grammatical error correction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 435–445, Florence, Italy. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2018. Inherent biases in reference-based evaluation for grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Melbourne, Australia. Association for Computational Linguistics.

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.

Simon Flachs, Ophélie Lacroix, Helen Yannakoudakis, Marek Rei, and Anders Søgaard. 2020. Grammatical error correction in low error density domains: A new benchmark and analyses. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8467–8478, Online. Association for Computational Linguistics.

Sylviane Granger. 1998. The computer learner corpus: A versatile new source of data for SLA research. *Learner English on Computer*, pages 3–18.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics.

Ola Knutsson, Teresa Cerrato Pargman, and Kerstin Severinson Eklundh. 2003. Transforming grammar checking technology into a learning environment for second language writing. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, pages 38–45.

Masato Mita and Hitomi Yanaka. 2021. Do grammatical error correction models realize grammatical generalization? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4554–4561, Online. Association for Computational Linguistics.

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.

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, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, 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.

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 the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. 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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering

challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2021. How good (really) are grammatical error correction systems? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2686–2698, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, pages 3261–3275.

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.

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.

Zheng Yuan and Christopher Bryant. 2021. Document-level grammatical error correction. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 75–84, Online. Association for Computational Linguistics.