Grammatical Error Correction in Low Error Density Domains: A New Benchmark and Analyses

Simon Flachs[1,2], **Ophélie Lacroix**[3]*,
Helen Yannakoudakis[4], **Marek Rei**[5], **Anders Søgaard**[2]
[1] Siteimprove sfl@siteimprove.com
[2] University of Copenhagen soegaard@di.ku.dk
[3] Alexandra Institute ophelie.lacroix@alexandra.dk
[4] King's College London helen.yannakoudakis@kcl.ac.uk
[5] Imperial College London marek.rei@imperial.ac.uk

Abstract

Evaluation of grammatical error correction (GEC) systems has primarily focused on essays written by non-native learners of English, which however is only part of the full spectrum of GEC applications. We aim to broaden the target domain of GEC and release CWEB, a new benchmark for GEC consisting of website text generated by English speakers of varying levels of proficiency. Website data is a common and important domain that contains far fewer grammatical errors than learner essays, which we show presents a challenge to state-of-the-art GEC systems. We demonstrate that a factor behind this is the inability of systems to rely on a strong internal language model in low error density domains. We hope this work shall facilitate the development of open-domain GEC models that generalize to different topics and genres.

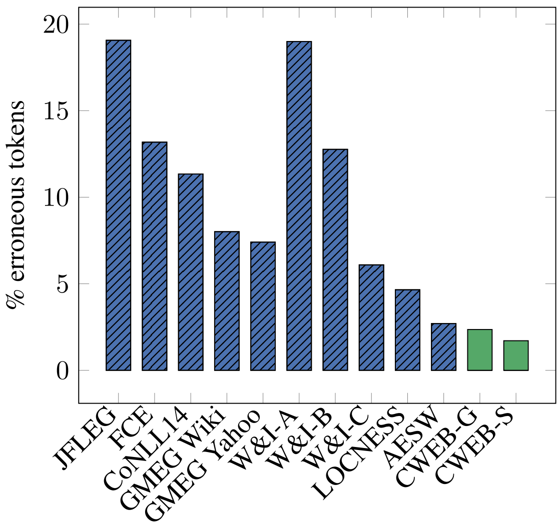


Figure 1: Percentage of erroneous tokens per domain. CWEB-G/S are our newly devised datasets.

1 Introduction

Grammatical error correction (GEC) is the task of automatically editing text to remove grammatical errors; for example: [*A link to registration can also be found ~~at~~* ***on*** *the same page.*]. GEC systems so far have primarily focused on correcting essays produced by English-as-a-second-language (ESL) learners, providing fast and inexpensive feedback to facilitate language learning. However, this is only one target domain in the full spectrum of GEC applications. GEC models can also help to improve written communication outside of the formal education setting. Today the largest medium of written communication is the internet, with approximately 380 new websites created every minute.[1] Ensuring grammatical correctness of websites helps facilitate clear communication and a professional commercial presentation. Therefore, it is important that GEC models perform well in the open-domain setting and generalize, not only to writing produced in the educational context, but also to language production "in the wild". Website data specifically represent a broad and diverse range of writing and constitute a major part of what people read and write on an everyday basis.

This work highlights two major prevailing challenges of current approaches to GEC: *domain adaptation* and *low precision* in texts with low error density. Previous work has primarily targeted essay-style text with high error density (see Figure 1); however, this lack of diversity means that it is not clear how systems perform on other domains and under different error distributions (Sakaguchi et al., 2017).[2]

Current publicly available datasets are restricted to non-native English essays [e.g. FCE (Yannakoudakis et al., 2011); CoNLL14 (Ng et al.,

[*] Research conducted at Siteimprove.

[1] https://www.millforbusiness.com/how-many-websites-are-there/

[2] Leacock et al. (2010) highlighted the variations in the distribution of errors in non-native and native English writings.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 8467–8478,
November 16–20, 2020. ©2020 Association for Computational Linguistics

Error type	Example sentence
VERB:SVA	They develop positive relationships with swimmers and members, and ~~promotes~~ **promote** programs in order to generate more participation.
MORPH / ORTH	In a small ~~agriculture~~ **agricultural** town on the east side of Washington ~~state~~ **State** called Yakima.
PREP	[. . .] the distance between the two should be ~~on~~ **of** the order of 50 microns.

Table 1: Example sentences from the CWEB dataset. Erroneous text is struck through and corrections are in bold.

2014)], student essays [W&I+LOCNESS (Bryant et al., 2019; Granger, 1998)] or target a specific domain [scientific writing; AESW (Daudaravicius et al., 2016)]. Supervised systems trained on specific domains are less likely to be as effective at correcting distinctive errors from other domains, as is the case for systems trained on learner data with different native languages (Chollampatt et al., 2016; Nadejde and Tetreault, 2019). The recent BEA 2019 shared task (Bryant et al., 2019) encouraged research in the use of low-resource and unsupervised approaches; however, evaluation primarily targeted the restricted domain of student essays. We show that when applied to data outside of the language learning domain, current state-of-the-art systems exhibit low precision due to a tendency to over-predict errors. Recent work tackled the domain adaptation problem, and released GEC benchmarks from Wikipedia data and online comments [GMEG Wiki+Yahoo (Napoles et al., 2019)]. However, these datasets present a high density of errors and represent a limited subset of the full distribution of errors in online writing.

Contributions: We (i) release a new dataset, CWEB (**C**orrected **Web**sites), of website data that is corrected for grammatical errors;[3] (ii) systematically compare it to previously released GEC corpora; (iii) benchmark current state-of-the-art GEC approaches on this data and demonstrate that they are heavily biased towards existing datasets with high error density, even after fine-tuning on our target domain; (iv) perform an analysis showing that a factor behind the performance drop is the inability of systems to rely on a strong internal language model in low error density domains.

We hope that the new dataset will contribute towards the development of robust GEC models in the open-domain setting.

		CWEB-S	CWEB-G	Total
Dev	sent.	2,862	3,867	6,729
	tokens	68,857	79,689	148,546
	edits	895	1595	2490
Test	sent.	2,864	3,981	6,845
	tokens	68,459	80,684	149,143
	edits	1004	1679	2683
Total	sent.	5,726	7,848	13,574
	tokens	137,316	160,373	297,689
	websites	453	625	1,078
	parag.	659	630	1,289

Table 2: Distribution of sentences and tokens in the CWEB dataset.

2 CWEB Dataset

We create a new dataset of English texts from randomly sampled websites, and annotate it for grammatical errors. The source texts are randomly selected from the first 18 dumps of the CommonCrawl[4] dataset and represent a wide range of data seen online such as blogs, magazines, corporate or educational websites. These include texts written by native or non-native English speakers and professional as well as amateur online writers.

Text Extraction To ensure English content, we exclude websites with country-code top-level domains; e.g., .fr, .de. We use the `jusText`[5] tool to retrieve the content from HTML pages (removing boilerplate elements and splitting the content into paragraphs). We heavily filter the data by removing paragraphs which contain non-English[6] and incomplete sentences. To ensure diversity of the data, we also remove duplicate sentences. Among the million sentences gathered, we select paragraphs randomly.

We split the data with respect to where they

[3] https://github.com/SimonHFL/CWEB

[4] https://commoncrawl.org/

[5] https://github.com/miso-belica/jusText

[6] Using the `langdetect` package.

	# sents	type-token	tok/sent	err. sents (%)	edits/sent	# annotators	sent-$\mathcal{K}$	NEs/sents
JFLEG	747	0.44	18.9	86.4	3.6	4	0.53	0.35
FCE	2,695	0.39	15.6	67.8	2.6	1	-†	0.59
CoNLL14	1,312	0.39	22.9	75.8	2.7	2	0.25	0.31
W&I-A	1,036	0.43	18.0	80.5	3.6	1	-†	0.58
W&I-B	1,285	0.45	18.4	72.1	2.7	1	-†	0.52
W&I-C	1,068	0.47	20.1	53.8	1.9	1	-†	0.78
LOCNESS	988	0.47	23.4	52.2	1.8	1	-†	0.77
GMEG wiki	992	0.55	26.9	82.3	2.5	4	0.43	2.83
GMEG yahoo	1,000	0.46	16.9	50.5	2.7	4	0.51	0.59
AESW	52,124	0.52	23.9	36.1	1.6	1	-†	0.93
CWEB-S	2,864	0.56	23.9	24.5	1.5	2	0.39	1.44
CWEB-G	3,981	0.53	20.3	25.6	1.9	2	0.44	1.04

Table 3: Statistics on GEC Corpora; type–token is the average ratio of vocabulary size by the total number of tokens (calculated as an average over a sliding window of $1,000$ tokens); ratio of edits per sentence is calculated on erroneous sentences; sent-$\mathcal{K}$ is sentence-level Cohen's Kappa score (†: calculated for datasets with > 1 annotator); NEs stands for Named Entities (extracted using Spacy).

come from: sponsored[7] (CWEB-S) or generic[8] (CWEB-G) websites. The sponsored data represent a more focused domain (professional writing) than the generic one which includes writing from various proficiency levels.

Annotation The data is corrected for errors by two expert annotators, trained for correcting grammatical errors in English text: not attempting to rewrite the text nor make fluency edits, but rather to make minimal edits – minimum number of edits to make the text grammatical. During error annotation, the annotators have access to the entire paragraph in which a sentence belongs, therefore using the context of a sentence to help them in the correction. Examples of erroneous sentences from our data are shown in Table 1. Annotator agreement is calculated at the sentence level using Cohen's Kappa, i.e. we calculate whether annotators agree on which sentences are erroneous. This approach is preferable to relying on exact matching of error corrections, as as there are often many different ways to correct a sentence (Bryant and Ng, 2015). Kappa is 0.39 and 0.44 for sponsored (CWEB-S) and generic website (CWEB-G) data respectively, and Table 3 presents how our agreement results compare to those of existing GEC datasets. The table also includes a number of other statistics, and the different datasets are further analyzed, compared and contrasted in Section 5.

The texts are tokenized using SpaCy[9] and automatically labeled for error types (and converted into the M2 format) using the ERRor ANnotation Toolkit (ERRANT) (Bryant et al., 2017).

Release For each dataset, we release a development and a test set: we propose a roughly equal division of the data into the two splits, which presents a fair amount of errors to evaluate on (see Table 2).

To avoid copyright restrictions, we split the collected paragraphs into sentences and shuffle all sentences in order to break the original and coherent structure that would be needed to reproduce the copyrighted material. This approach has successfully been used in previous work for devising web-based corpora (Schäfer, 2015; Biemann et al., 2007). The data is available at `https://github.com/SimonHFL/CWEB`.

3 GEC Corpora

We compare our data with existing GEC corpora which cover a range of domains and proficiency levels. Table 3 presents a number of different statistics and Table 4 their error-type frequencies.[10]

3.1 English as a second language (ESL)

JFLEG (Napoles et al., 2017) The JHU Fluency-Extended GUG corpus consists of sentences written by English language learners (with different proficiency levels and L1s) for the TOEFL® exam,

[7] top-level domains: .gov, .edu, .mil, .int, and .museum.

[8] top-level domains: .com, .info, .net, .org.

[9] `https://spacy.io/`

[10] See links to downloadable versions in Appendix A

	JFLEG	FCE 2.1	CoNLL14	W&I			LOCNESS	GMEG		AESW	CWEB	
				A	B	C		Wiki	Yahoo		G	S
Punct	147.7	112.3	65.5	244.8	188.2	100.4	152.3	230.0	194.0	80.6	48.9	48.7
Verb	233.5	176.7	200.5	300.0	202.5	79.4	19.9	48.1	24.2	17.8	23.4	13.1
Other	295.6	138.3	158.1	237.3	136.7	57.4	43.3	93.8	98.0	42.7	31.6	21.0
Det	180.7	149.1	134.9	159.1	124.1	65.8	16.4	40.3	22.6	33.7	20.9	19.7
Noun	167.7	105.4	116.8	139.8	89.0	49.9	32.4	63.6	26.2	16.8	19.6	12.8
Prep	107.1	113.8	92.7	137.2	114.4	64.9	28.1	37.1	21.1	11.4	15.6	9.8
Spell	242.5	107.8	26.0	79.3	36.3	16.3	51.0	86.9	68.0	5.1	3.8	2.4
All	1675.6	1084.9	919.6	1561.2	1050.7	504.1	400.6	732.3	635.3	239.2	208.9	147.2

Table 4: Number of error occurrences for the most frequent error types (per $10,000$ token).

covering a range of topics. Texts have been corrected for grammatical errors and fluency.

FCE (Yannakoudakis et al., 2011) consists of $1,244$ error corrected texts produced by learners taking the First Certificate in English exam, which assesses English at an upper-intermediate level. We use the data split made available for the BEA GEC shared task 2019 (Bryant et al., 2019).

CoNLL14 (Ng et al., 2014) consists of (mostly argumentative) essays written by ESL learners from the National University of Singapore, which are annotated for grammatical errors by two native speakers of English.

Write&Improve (W&I) (Bryant et al., 2019) Cambridge English Write & Improve (Yannakoudakis et al., 2018) is an online web platform that automatically provides diagnostic feedback to non-native English-language learners, including an overall language proficiency score based on the Common European Framework of Reference for Languages (CEFR).[11] The W&I corpus contains $3,600$ texts across 3 different CEFR levels – A (beginner), B (intermediate), and C (advanced) – that have been annotated for errors.[12]

3.2 Other Corpora

LOCNESS (Bryant et al., 2019; Granger, 1998) The LOCNESS corpus consists of essays written by native English students. A sample of 100 essays has been annotated for errors with a 50:50 development/test split.[13]

GMEG Wiki (Napoles et al., 2019) is devised based on edits in the Wikipedia revision history, and the writing therefore represents formal articles. Note that collecting sentences based on edits in the Wikipedia revision history introduces a substantial bias.[14] This means that evaluation results on this benchmark are not truly representative of how a system would perform when applied to realistic online data and full-length articles.

GMEG Yahoo (Napoles et al., 2019) comprises paragraphs of informal web posts gathered from answers in the *Yahoo! Answers* platform. The style is informal, and contains slang terms and non-conventional mechanics.

AESW (Daudaravicius et al., 2016) was released as part of the Automated Evaluation of Scientific Writing Shared Task. It is a collection of text extracts from published journal articles (mostly in physics and mathematics) along with their (sentence-aligned) corrected counterparts.[15]

4 System Performance

We evaluate performance on GEC benchmarks for two approaches to GEC that currently have state-of-the-art performance on CoNLL14. The first approach, that we refer to as GEC-pseudodata and is proposed by Kiyono et al. (2019),[16] uses a transformer-based seq2seq model. The second approach uses the PIE system (Awasthi et al., 2019)[17] which leverages a BERT-based architecture for local sequence transduction tasks. Both models are

[11] `https://www.cambridgeenglish.org/exams-and-tests/cefr/`

[12] Since error corrections on test sets are not publicly available, we carry out our analyses on the development sets.

[13] See footnote 12.

[14] Sentences that have been edited are more likely to contain grammatical errors, and grammatical errors will therefore be over-represented. This is reflected in the 82.3% erroneous sentence rate (see Table 3).

[15] We exclude sentences that use AESW's normalization scheme (e.g. citations replaced with _CITE_), as the models we use are not trained with these special tokens.

[16] `www.github.com/butsugiri/gec-pseudodata`; We use the PRETLARGE+SSE (finetuned) model.

[17] `www.github.com/awasthiabhijeet/PIE`

	JFLEG	FCE	CoNLL14	W&I			LOCNESS	GMEG		AESW	CWEB		
				A	B	C		Wiki	Yahoo		G	S	G+S
GEC-PSEUDODATA system													
P	55.73	55.11	44.96	54.89	54.86	44.53	47.09	52.81	37.57	14.05	21.34	17.27	19.97
R	38.73	41.61	29.03	37.92	35.14	32.04	34.13	23.02	32.26	13.24	23.00	15.75	20.28
$F_{0.5}$	51.13	51.75	40.35	50.38	49.32	41.31	43.77	41.89	36.00	13.88	21.58	16.91	19.98
PIE system													
P	51.04	49.55	43.47	50.24	49.12	39.12	32.77	44.71	33.08	8.78	14.29	5.73	10.80
R	35.21	36.34	27.93	36.10	31.20	27.13	23.11	19.66	26.97	9.67	18.91	8.78	15.11
$F_{0.5}$	46.74	46.19	38.95	46.59	44.06	35.94	30.24	35.58	31.29	8.94	14.98	6.15	11.43

Table 5: Scores of two SOTA GEC systems on each domain. For both systems performance is substantially lower on CWEB than ESL domains. Scores are calculated against each individual annotator and averaged

pre-trained on synthetic errors and fine-tuned on learner data from the train section of FCE (Yannakoudakis et al., 2011), Lang-8 (Mizumoto et al., 2011), and NUCLE (Dahlmeier et al., 2013) and for GEC-PSEUDODATA additionally on the W&I train split (Bryant et al., 2019).

Performance is evaluated using the $F_{0.5}$ metric calculated by ERRANT (Bryant et al., 2017).[18] However, the more annotators a dataset has, the higher score a system will get on this data (Bryant and Ng, 2015). In order to perform a fair comparison of systems across datasets with a different number of annotators, we calculate the ERRANT score against each individual annotator and then take the average to get the final score.

Evaluation results are presented in Table 5. Across all datasets, we observe lower scores with the PIE system (-6.05 $F_{0.5}$ on average), while GEC-PSEUDODATA is consistently better. Overall $F_{0.5}$ ranges from around 30 to 52 for most datasets; however, when the models are evaluated on CWEB and AESW, we observe a substantial drop in performance, with the lowest $F_{0.5}$ score being the PIE system on CWEB-S (6.15). Precision, in particular, suffers due to the systems over-correcting sentences that should remain unchanged.

Using the GEC-PSEUDODATA system, on average, we find a higher $F_{0.5}$ on ESL corpora (JFLEG, FCE, CoNLL, W&I) compared to non-ESL ones (47.4 vs. 29.0). This demonstrates that GEC systems trained on language learning data do not perform as well on other domains and further work is needed to improve their generalization.

	P	R	$F_{0.5}$
CWEB-G	42.09	16.56	32.01
CWEB-S	35.91	12.96	26.46
CWEB (G+S)	39.89	15.2	30.0

Table 6: Scores of the GEC-PSEUDODATA system fine-tuned on CWEB data. Fine-tuning yields substantial improvements, but scores are still worse than on ESL domains. Scores are calculated against each individual annotator and averaged.

4.1 Fine-tuning

We investigate the extent to which the GEC-PSEUDODATA system can be adapted to our domain, and fine-tune it using our development sets.[19] We take $1,000$ sentences from each of the development sets of CWEB-G and CWEB-S and use them as a development set for this experiment. The remaining $4,729$ sentences of our development sets are used as training data for fine-tuning the GEC system.

In Table 6, we can see that fine-tuning substantially improves performance (around $+10.0$ $F_{0.5}$ across all CWEB sets). In particular, precision is improved ($+20.8$/$+18.6$ on CWEB-G/S) at the expense of recall (-6.4/-2.8 on CWEB-G/S). However, performance is still low compared to the language learning domain ($F_{0.5}$ of at least 41), further indicating that there is scope for developing more robust and general-purpose, open-domain GEC systems. For the purpose of future benchmarking, Appendix B lists the system's ERRANT scores based on both annotators – as opposed to the average of individual annotator scores reported in Table 6.

[18] www.github.com/chrisjbryant/errant

[19] We use the fine-tuning parameters of Kiyono et al. (2019).

5 Analysis

In order to assess the impact our new dataset can have on the GEC field, we carry out analyses to show 1) to what degree the domain of our data is different from existing GEC corpora, and how existing GEC systems are affected by the domain shift; and 2) that a factor behind the performance drop on CWEB data is the inability of systems to rely on a strong internal language model in low error density domains.

5.1 Domain Shift

Moving from error correction in learner texts to error correction in diverse, online texts, many of which are written by professional writers, amounts to a drift in data distribution. In general, distributional drift comes in different flavors; given two distributions $P(\mathbf{X}, \mathbf{Y})$ and $Q(\mathbf{X}, \mathbf{Y})$:

Covariate shift concerns change in the marginal distribution of the independent variable, i.e., $P(\mathbf{X}) \neq Q(\mathbf{X})$. In the context of grammatical errors, this refers to the degree to which the type of sentences written varies between domains. Table 3 clearly shows covariate shift effects: see, for example, differences in vocabulary variation (measured by the type–token ratio) and the frequency of named entities.

Label bias describes the change in distribution of the dependent variable, i.e., $P(\mathbf{Y}) \neq Q(\mathbf{Y})$. In terms of GEC, this refers to the difference in error distributions across domains. In Table 3, we can see that CWEB data contains errors that are substantially more sparse than other domains – a smaller proportion of sentences are erroneous, and these erroneous sentences also contain fewer edits compared to other domains. Additionally, looking at Table 4, we can see that almost all error types are substantially less frequent in our data than in existing benchmarks – for example, spelling errors are 38 times more prevalent in GMEG Wiki compared to CWEB-S.

Moving from learner text to web data involves both forms of drift: covariate shift and label bias. We further analyze the effects of these shifts on system performance.

5.1.1 Impact of Error Density

To demonstrate that the error density of corpora has a substantial impact on the performance of GEC systems, we vary the proportion of erroneous sentences in each dataset by either removing correct sentences or by adding correct sentences of the same domain.[20] By fixing the frequency of errors across datasets, we can observe, in isolation, how the systems are affected by co-variate shift across domains. Precision as a function of the proportion of erroneous sentences for selected datasets[21] is presented in Figure 2 (recall is unchanged).

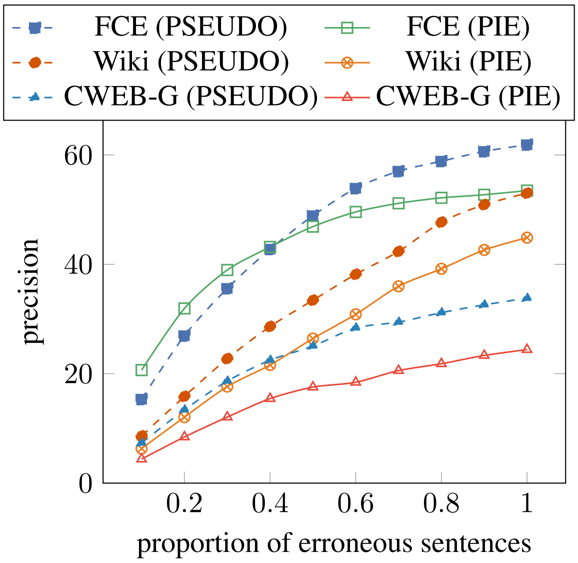


Figure 2: Precision as a function of the proportion of erroneous sentences in 3 different domains; comparing the GEC-PSEUDODATA (PSEUDO) and PIE systems.

For each domain, we observe precision being highly sensitive to the proportion of errors. This indicates that differences in error distribution across domains (i.e. label bias) is likely to be a large contributor to performance drop. We also observe the effect of covariate shift across the datasets: while the percentage of erroneous sentences is the same, precision differs for the different datasets which suggests that covariate shift across domains has an impact on the performance of the system.

5.1.2 Analysis of Gold Edits

In addition to error density, the type of errors present in the dataset also has an impact on the performance of GEC systems. We investigate how errors and their corresponding corrections differ across domains. In particular, we look at how gold edits in different domains change the sentence in terms of two factors: 1) How much do edits change the semantics of the sentence, and 2) to what degree do edits improve the sentence.

[20] For each dataset, we apply the gold corrections on incorrect sentences, creating new examples of in-domain, correct sentences, which are then randomly selected for inclusion.

[21] Scores for all datasets can be found in Appendix D.

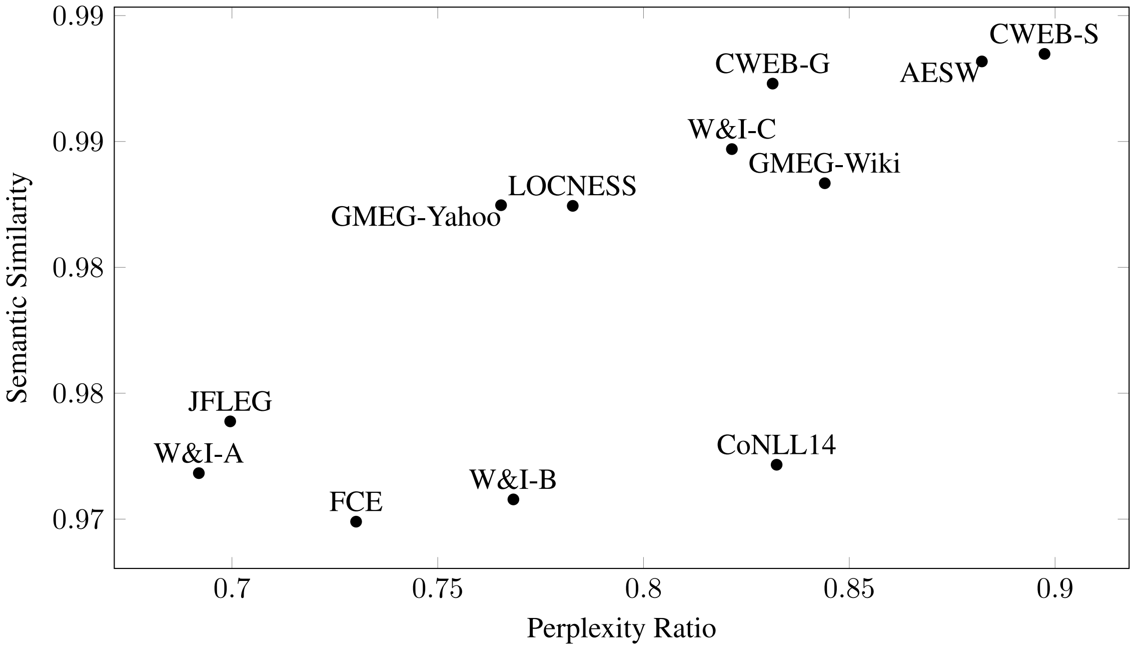


Figure 3: Average semantic similarity and perplexity ratio (sentence improvement) of sentences before and after being edited, plotted per dataset. The analysis is limited to sentences containing exactly one edit.

We limit our analysis to sentences containing exactly one edit, as we are interested in how individual edits change a sentence, regardless of how domains differ in amounts of erroneous sentences and in the number of edits per sentence (Table 3).

Regarding 1), to measure the semantic change of a sentence after an edit is introduced, we use sentence embeddings generated by Sentence-BERT (Devlin et al., 2019) and calculate the cosine similarity between the original sentence and its corrected counterpart. Regarding 2), the degree of sentence improvement is calculated as the ratio of the perplexity of GPT-2 (Radford et al., 2019) on a sentence after and before it has been edited.

$$\Delta P = \frac{PPL(edited\_sentence)}{PPL(original\_sentence)}$$

A lower ratio suggests that the edited sentence is an improvement, since its perplexity is lower than the original sentence.

Using the outputs of machine learning models as a proxy for semantic change and sentence improvement inevitably introduces biases, but nevertheless provide valuable insights into domain differences.

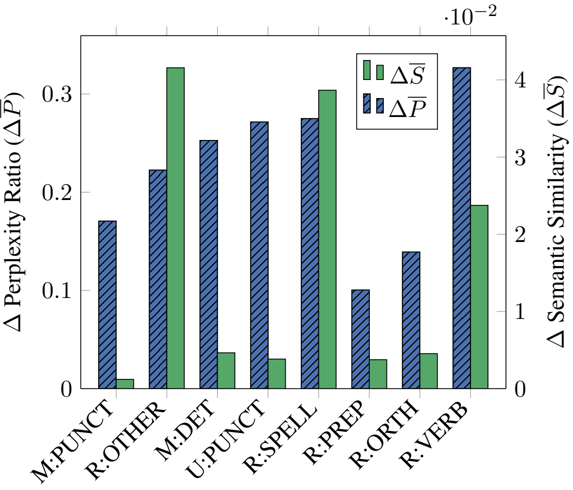


Figure 4: Difference in semantic similarity and perplexity ratio between CWEB-S and FCE for the most frequent error types (M: missing; R: replace; U: unnecessary).

Corpus Level In Figure 3, the average semantic similarity and perplexity ratio is plotted for each dataset. It is evident that ESL datasets consist of edits with a higher degree of semantic change and sentence improvements than datasets from more advanced speakers. CWEB and AESW in particular stand out, with edits that largely retain the semantics of a sentence and that result in more subtle improvements.

Error type level In order to gain further insight on what is driving the differences between datasets, we look separately at how edits of each error type change the sentence. We compare FCE and CWEB-S, which lie at opposite ends in Figure 3. For each

	P	R	F$_{0.5}$
JFLEG	57.55	21.59	43.07
FCE	51.33	17.39	36.92
CoNLL14	40.30	16.56	31.17
W&I-A	45.79	15.10	32.55
W&I-B	43.17	14.46	30.90
W&I-C	33.02	9.81	22.42
LOCNESS	42.09	16.09	31.81
GMEG Wiki	52.36	13.35	32.99
GMEG Yahoo	62.50	16.45	39.45
AESW	10.18	3.58	7.44
CWEB-G	15.20	5.96	11.54
CWEB-S	8.94	1.33	4.17

Table 7: Scores of a language model based GEC system. The lower scores on CWEB and AESW indicate an inability to rely on language modelling in low error-density domains.

False Positive Examples	Perplexity ratio
All types of work are ~~callings~~ **called** to individuals.	0.34
Get started ~~at~~ **with** ACC	0.51
That ~~is~~ **was** actually kind of fun!	0.69

Table 8: Examples of false positives on the CWEB dataset that improve perplexity substantially – even more than the average gold edit in CWEB (0.86 perplexity ratio).

dataset, we obtain an average of semantic similarity, $\overline{S}$, and perplexity ratio, $\overline{P}$, separately for sentences of each error type. Then, for each error type, the difference, Δ, between scores in the two datasets is calculated.

$$\Delta\overline{S} = \overline{S}_{\text{CWEB-S}} - \overline{S}_{\text{FCE}}$$

$$\Delta\overline{P} = \overline{P}_{\text{CWEB-S}} - \overline{P}_{\text{FCE}}$$

Figure 4 plots these differences for the most common error types. We can observe that, for all error types, edits in CWEB-S result in both a lower degree of semantic change and sentence improvement than edits in FCE. This is particularly evident for the error types R:OTHER, R:SPELL and R:VERB. These are open class errors, where the error and correction can be quite different. It is therefore reasonable that differences in edits' degree of semantic change and perplexity improvement across domains are particularly observed in these cases.[22]

5.2 Language Model Importance

We also investigate the degree to which systems can rely on a strong internal language model representation when evaluated against different domains. We examine this by looking at the performance of a purely language model based GEC system over the different datasets.

We build on the approach of Bryant and Briscoe (2018), using confusion sets to generate alternative versions of an input sentence and then deciding if any of the alternatives are preferable to the original version, based on language model probabilities. The authors use an n-gram language model, which we replace with GPT-2 (Radford et al., 2019) to see how a strong neural language model performs – this approach is similar to Alikaniotis and Raheja (2019). Hyperparameters are tuned for each dataset (see Appendix C for details).

Table 7 displays the results on the different datasets. Recall and, in particular, precision is substantially lower on CWEB and AESW compared to other datasets. In general, scores are higher in domains with a higher proportion of errors and those containing edits which result in high perplexity improvements. In these cases systems can rely on a rough heuristic of replacing low probability sequences with high probability ones. However, in CWEB, where errors are fewer and more subtle, this leads to low precision, as perplexity alone cannot differentiate an erroneous sequence from a sequence that is rare but correct. Table 8 displays several examples of this, where false positive corrections suggested by the language model based GEC system have large perplexity improvements.

This analysis suggests that the inability to rely on a strong internal language model representation can negatively impact SOTA system performance on CWEB and on low error density domains in general. This would mean that having large amounts of error examples for training is more important in high-level domains.

6 Conclusion

We release a new GEC benchmark, CWEB, consisting of website text generated by English speakers at varying levels of proficiency. Comparisons against existing benchmarks demonstrate that CWEB differs in many respects: 1) in the distribution of sentences (higher vocabulary variation and named entity frequency); 2) in error density (lower); and 3)

[22] Score differences for the R:SPELL error type seem to be driven by a different propensity of spelling errors being of a typographical vs. phonetical nature in the two datasets.

in the types of edits and their impact on language model perplexity and semantic change.

We showed that existing state-of-the-art GEC models achieve considerably lower performance when evaluated on this new domain, even after finetuning. We argue that a factor behind this is the inability of systems to rely on a strong internal language model in low error density domains.

We hope that the dataset shall broaden the target domain of GEC beyond learner and/or exam writing and facilitate the development of robust GEC models in the open-domain setting.

References

Dimitrios Alikaniotis and Vipul Raheja. 2019. The Unreasonable Effectiveness of Transformer Language Models in Grammatical Error Correction. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA 2019. Association for Computational Linguistics.

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel Iterative Edit Models for Local Sequence Transduction. 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 2019. Association for Computational Linguistics.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig Corpora Collection: Monolingual Corpora of Standard Size. In *Proceedings of the Corpus Linguistics Conference*, CL2007.

Christopher Bryant and Ted Briscoe. 2018. Language Model Based Grammatical Error Correction without Annotated Training Data. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA 2018. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Edward John 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*, ACL 2017. 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*, ACL-IJCNLP 2015. 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*, EMNLP 2016. 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 8th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA 2013. Association for Computational Linguistics.

Vidas Daudaravicius, Rafael E. Banchs, Elena Volodina, and Courtney Napoles. 2016. A Report on the Automatic Evaluation of Scientific Writing Shared Task. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA 2016. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL 2019. Association for Computational Linguistics.

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

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 2019. Association for Computational Linguistics.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. Automated Grammatical Error Detection for Language Learners. *Synthesis lectures on human language technologies*, 3(1):1–134.

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 2011*.

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. Association for Computational Linguistics.

Courtney Napoles, Maria Nădejde, and Joel Tetreault. 2019. Enabling Robust Grammatical Error Correction in New Domains: Data Sets, Metrics, and Analyses. *Transactions of the Association for Computational Linguistics (TACL 2019)*, 7:551–566.

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*, EACL 2017. 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 18th Conference on Computational Natural Language Learning: Shared Task*, CoNLL 2014. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Keisuke Sakaguchi, Courtney Napoles, and Joel Tetreault. 2017. GEC into the future: Where are we going and how do we get there? In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, CMLC-3. Institut für Deutsche Sprache.

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*, ACL 2011. Association for Computational Linguistics.

A Dataset Download Links

- JFLEG: https://github.com/keisks/jfleg
- FCE: https://www.cl.cam.ac.uk/research/nl/bea2019st/#data
- CoNLL14: https://www.comp.nus.edu.sg/~nlp/conll14st.html
- Write&Improve-A/B/C: https://www.cl.cam.ac.uk/research/nl/bea2019st/#data
- LOCNESS: https://www.cl.cam.ac.uk/research/nl/bea2019st/#data
- GMEG Yahoo/Wiki: https://github.com/grammarly/GMEG
- AESW: http://textmining.lt/aesw/aesw2016down.html

B Non-averaged Fine-tuning Scores

	P	**R**	**F$_{0.5}$**
CWEB-G	53.88	34.24	48.33
CWEB-S	43.65	31.1	40.39
CWEB (all)	50.25	33.2	45.57

Table 9: Scores of the GEC-PSEUDODATA system fine-tuned on CWEB data, calculated against both annotators.

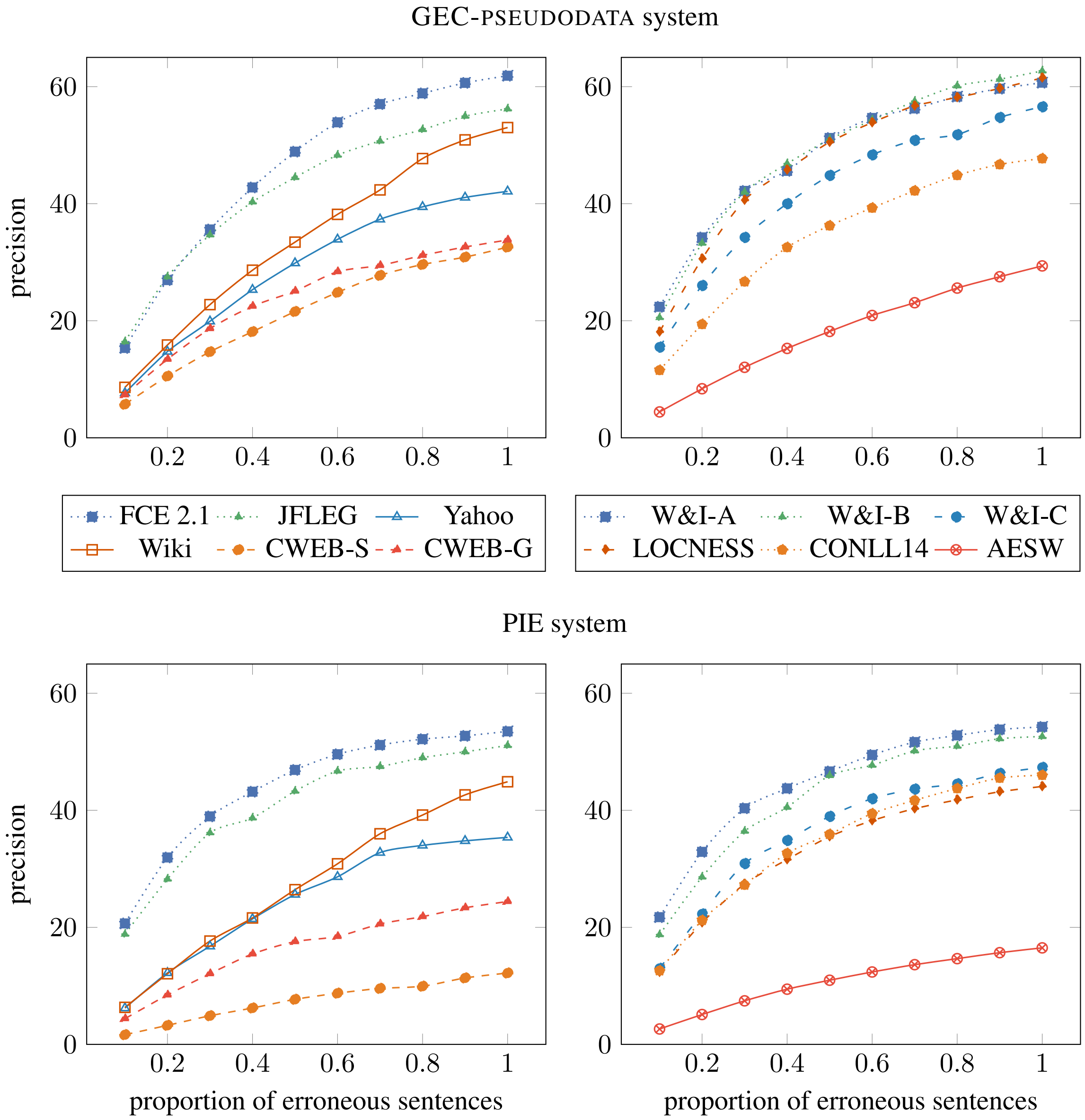
C Language Model GEC Hyperparameter Tuning

A threshold, τ, determines the degree of probability improvement needed before an alternative sentence is preferred. For each dataset, we find τ, in the 0.9 to 1.0 range, resulting in the best development set $F_{0.5}$. For CoNLL14, we tune on CoNLL13; for W&I, we use the dedicated training sets; for LOCNESS, there is no training set available and so we tune on the W&I subset of advanced texts (W&I-C).

	JFLEG	**FCE**	**CoNLL14**	**W&I**			**LOCNESS**	**GMEG**		**AESW**	**CWEB**	
				A	**B**	**C**		**Wiki**	**Yahoo**		**G**	**S**
τ	0.97	0.97	0.98	0.98	0.98	0.97	0.97	0.96	0.91	0.96	0.96	0.93

Table 10: Best performing threshold τ for each domain.

D Precision as a Function of the Proportion of Erroneous Sentences



Precision as a function of the proportion of erroneous sentences in each domain.