JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction

Courtney Napoles,[1] **Keisuke Sakaguchi**,[1] and **Joel Tetreault**[2]

[1]Center for Language and Speech Processing, Johns Hopkins University

[2]Grammarly

{napoles,keisuke}@cs.jhu.edu, joel.tetreault@grammarly.com

Abstract

We present a new parallel corpus, **J**HU **FL**uency-**E**xtended **G**UG corpus (JFLEG) for developing and evaluating grammatical error correction (GEC). Unlike other corpora, it represents a broad range of language proficiency levels and uses holistic *fluency edits* to not only correct grammatical errors but also make the original text more native sounding. We describe the types of corrections made and benchmark four leading GEC systems on this corpus, identifying specific areas in which they do well and how they can improve. JFLEG fulfills the need for a new gold standard to properly assess the current state of GEC.

1 Introduction

Automatic grammatical error correction (GEC) progress is limited by the corpora available for developing and evaluating systems. Following the release of the test set of the CoNLL–2014 Shared Task on GEC (Ng et al., 2014), systems have been compared and new evaluation techniques proposed on this single dataset. This corpus has enabled substantial advancement in GEC beyond the shared tasks, but we are concerned that the field is over-developing on this dataset. This is problematic for two reasons: 1) it represents one specific population of language learners; and 2) the corpus only contains *minimal edits*, which correct the grammaticality of a sentence but do not necessarily make it *fluent* or native-sounding.

To illustrate the need for fluency edits, consider the example in Table 1. The correction with only minimal edits is grammatical but sounds *awkward* (unnatural to native speakers). The fluency correction has more extensive changes beyond addressing grammaticality, and the resulting sen-

Original: they just creat impression such well that people are drag to buy it .
Minimal edit: They just create an impression so well that people are dragged to buy it .
Fluency edit: They just create such a good impression that people are compelled to buy it.

Table 1: A sentence corrected with just minimal edits compared to fluency edits.

tence sounds more natural and its intended meaning is more clear. It is not unrealistic to expect these changes from automatic GEC: the current best systems use machine translation (MT) and are therefore capable of making broader sentential rewrites but, until now, there has not been a gold standard against which they could be evaluated.

Following the recommendations of Sakaguchi et al. (2016), we release a new corpus for GEC, the **J**HU **FL**uency-**E**xtended **G**UG corpus (JFLEG), which adds a layer of annotation to the GUG corpus (Heilman et al., 2014). GUG represents a cross-section of ungrammatical data, containing sentences written by English language learners with different L1s and proficiency levels. For each of 1,511 GUG sentences, we have collected four human-written corrections which contain holistic fluency rewrites instead of just minimal edits. This corpus represents the diversity of edits that GEC needs to handle and sets a gold standard to which the field should aim. We overview the current state of GEC by evaluating the performance of four leading systems on this new dataset. We analyze the edits made in JFLEG and summarize which types of changes the systems successfully make, and which they need to address. JFLEG will enable the field to move beyond minimal error corrections.

2 GEC corpora

There are four publicly available corpora of non-native English annotated with corrections, to our

Corpus	# sents.	Mean chars per sent.	Sents. changed	Mean LD
AESW	1.2M	133	39%	3
FCE	34k	74	62%	6
Lang-8	1M	56	35%	4
NUCLE	57k	115	38%	6
JFLEG	1,511	94	86%	13

Table 2: Parallel corpora available for GEC.

knowledge. The NUS Corpus of Learner English (NUCLE) contains essays written by students at the National University of Singapore, corrected by two annotators using 27 error codes (Dahlmeier et al., 2013). The CoNLL Shared Tasks used this data (Ng et al., 2014; Ng et al., 2013), and the 1,312 sentence test set from the 2014 task has become *de rigueur* for benchmarking GEC. This test set has been augmented with ten additional annotations from Bryant et al. (2015) and eight from Sakaguchi et al. (2016). The Cambridge Learner Corpus First Certificate in English (FCE) has essays coded by one rater using about 80 error types, alongside the score and demographic information (Yannakoudakis et al., 2011). The Lang-8 corpus of learner English is the largest, with text from the social platform lang-8.com automatically aligned to user-provided corrections (Tajiri et al., 2012). Unlimited annotations are allowed per sentence, but 87% were corrected once and 12% twice. The AESW 2016 Shared Task corpus contains text from scientific journals corrected by a single editor. To our knowledge, AESW is the only corpus that has not been used to develop a GEC system.

We consider NUCLE[1] and FCE to contain *minimal edits*, since the edits were constrained by error codes, and the others to contain *fluency* edits since there were no such restrictions. English proficiency levels vary across corpora: FCE and NUCLE texts were written by English language learners with relatively high proficiency, but Lang-8 is open to any internet user. AESW has technical writing by the most highly proficient English writers. Roughly the same percent of sentences from each corpus is corrected, except for FCE which has significantly more. This may be due to the rigor of the annotators and not the writing quality.

The following section introduces the JFLEG corpus, which represents a diversity of potential corrections with four corrections of each sentence. Unlike NUCLE and FCE, JFLEG does not restrict corrections to minimal error spans, nor are the er-

[1]Not including the additional fluency edits collected for the CoNLL-2014 test set by Sakaguchi et al. (2016).

rors coded. Instead, it contains holistic sentence rewrites, similar to Lang-8 and AESW, but contains more reliable corrections than Lang-8 due to perfect alignments and screened editors, and more extensive corrections than AESW, which contains fewer edits than the other corpora with a mean Levenshtein distance (LD) of 3 characters. Table 2 provides descriptive statistics for the available corpora. JFLEG is also the only corpus that provides corrections alongside sentence-level grammaticality scores of the uncorrected text.

3 The JFLEG corpus

Our goal in this work is to create a corpus of fluency edits, following the recommendations of (Sakaguchi et al., 2016), who identify the shortfalls of minimal edits: they artificially restrict the types of changes that can be made to a sentence and do not reflect the types of changes required for native speakers to find sentences *fluent*, or natural sounding. We collected annotations on a public corpus of ungrammatical text, the GUG (Grammatical/Ungrammatical) corpus (Heilman et al., 2014). GUG contains $3.1k$ sentences written by English language learners for the TOEFL® exam, covering a range of topics. The original GUG corpus is annotated with grammaticality judgments for each sentence, ranging from 1–4, where 4 is perfect or native sounding, and 1 incomprehensible. The sentences were coded by five crowd-sourced workers and one expert. We refer to the mean grammaticality judgment of each sentence from the original corpus as the *GUG score*.

In our extension, JFLEG, the 1,511 sentences which comprise the GUG development and test sets were corrected four times each on Amazon Mechanical Turk. Annotation instructions are included in Table 3. 50 participants from the United States passed a qualifying task of correcting five sentences, which was reviewed by the authors (two native and one proficient non-native speakers of American English). Annotators also rated how difficult it was for them to correct each sentence on a 5-level Likert scale (5 being very easy and 1 very difficult). On average, the sentences were relatively facile to correct (mean difficulty of 3.5 ± 1.3), which moderately correlates with the GUG score (Pearson's $r = 0.47$), indicating that less grammatical sentences were generally more difficult to correct. To create a blind test set for the community, we withhold half (747) of the sentences from the analysis and evaluation herein.

> Please correct the following sentence to make it sound natural and fluent to a native speaker of (American) English. The sentence is written by a second language learner of English. You should fix grammatical mistakes, awkward phrases, spelling errors, etc. following standard written usage conventions, but your edits must be conservative. Please keep the original sentence (words, phrases, and structure) as much as possible. The ultimate goal of this task is to make the given sentence sound natural to native speakers of English without making unnecessary changes. Please do not split the original sentence into two or more. Edits are not required when the sentence is already grammatical and sounds natural.

Table 3: JFLEG annotation instructions.

		Error type in original		
		Awkward	**Ortho.**	**Grammatical**
Edit type	Fluency	38%	35%	32%
	Minimal	82%	89%	85%

Table 4: Percent of sentences by error type that were changed with fluency or minimal edits.

The mean LD between the original and corrected sentences is more than twice that of existing corpora (Table 2). LD negatively correlates with the GUG score ($r = -0.41$) and the annotation difficulty score (-0.37), supporting the intuition that less grammatical sentences require more extensive changes, and it is harder to make corrections involving more substantive edits. Because there is no clear way to quantify agreement between annotators, we compare the annotations of each sentence to each other. The mean LD between all pairs of annotations is greater than the mean LD between the original and corrected sentences (15 characters), however 36% of the sentences were corrected identically by at least two participants.

Next, the English L1 authors examined 100 randomly selected original and human-corrected sentence pairs and labeled them with the type of error present in the sentence and the type of edit(s) applied in the correction. The three error types are sounds *awkward* or has an *orthographic* or *grammatical* error.[2] The majority of the original sentences have at least one error (81%), and, for 68% of these sentences, the annotations are error free. Few annotated sentences have orthographic (4%) or grammatical (10%) errors, but awkward errors are more frequent (23% of annotations were labeled *awkward*)—which is not very surprising given how garbled some original sentences are and the dialectal variation of what sounds awkward.

The corrected sentences were also labeled with

[2]Due to their frequency, we separate orthographic errors (spelling and capitalization) from other grammatical errors.

the type of changes made (minimal and/or fluency edits). Minimal edits reflect a minor change to a small span (1–2 tokens) addressing an immediate grammatical error, such as number agreement, tense, or spelling. Fluency edits are more holistic and include reordering or rewriting a clause, and other changes that involve more than two contiguous tokens. 69% of annotations contain at least one minimal edit, 25% a fluency edit, and 17% both fluency and minimal edits. The distribution of edit types is fairly uniform across the error type present in the original sentence (Table 4). Notably, fewer than half of awkward sentences were corrected with fluency edits, which may explain why so many of the corrections were still *awkward*.

4 Evaluation

To assess the current state of GEC, we collected automated corrections of JFLEG from four leading GEC systems with no modifications. They take different approaches but all use some form of MT. The best system from the CoNLL-2014 Shared Task is a hybrid approach, combining a rule-based system with MT and language-model reranking (CAMB14; Felice et al., 2014). Other systems have been released since then and report improvements on the 2014 Shared Task. They include a neural MT model (CAMB16; Yuan and Briscoe, 2016), a phrase-based MT (PBMT) with sparse features (AMU; Junczys-Dowmunt and Grundkiewicz, 2016), and a hybrid system that incorporates a neural-net adaptation model into PBMT (NUS; Chollampatt et al., 2016).

We evaluate system output against the four sets of JFLEG corrections with GLEU, an automatic fluency metric specifically designed for this task (Napoles et al., 2015) and the Max-Match metric (M^2) (Dahlmeier and Ng, 2012). GLEU is based on the MT metric BLEU, and represents the n-gram overlap of the output with the human-corrected sentences, penalizing n-grams that were been changed in the human corrections but left unchanged by a system. It was developed to score fluency in addition to minimal edits since it does not require an alignment between the original and corrected sentences. M^2 was designed to score minimal edits and was used in the CoNLL 2013 and 2014 shared tasks on GEC (Ng et al., 2013; Ng et al., 2014). Its score is the $F_{0.5}$ measure of word and phrase-level changes calculated over a lattice of changes made between the aligned origi-

System	TrueSkill	GLEU	M^2	Sentences changed
Original	-1.64	38.2	0.0	–
CAMB16	**0.21**	**47.2**	50.8	74%
NUS	-0.20*	46.3	**52.7**	69%
AMU	-0.46*	41.7	43.2	56%
CAMB14	-0.51*	42.8	46.6	58%
Human	2.60	55.3	63.2	86%

Table 5: Scores of system outputs. * indicates no significant difference from each other.

nal and corrected sentences. Since both GLEU and M^2 have only been evaluated on the CoNLL-2014 test set, we additionally collected human rankings of the outputs to determine whether human judgments of relative grammaticality agree with the metric scores when the reference sentences have fluency edits.

The two native English-speaking authors ranked six versions of each of 150 JFLEG sentences: the four system outputs, one randomly selected human correction, and the original sentence. The absolute human ranking of systems was inferred using TrueSkill, which computes a relative score from pairwise comparisons, and we cluster systems with overlapping ranges into equivalence classes by bootstrap resampling (Sakaguchi et al., 2014; Herbrich et al., 2006). The two best ranked systems judged by humans correspond to the two best GLEU systems, but GLEU switches the order of the bottom two. The M^2 ranking does not perform as well, reversing the order of the top two systems and the bottom two (Table 5).[3] The upper bound is GLEU = 55.3 and M^2 = 63.2, the mean metric scores of each human correction compared to the other three. CAMB16 and NUS are halfway to the gold-standard performance measured by GLEU and, according to M^2, they achieve approximately 80% of the human performance. The neural methods (CAMB16 and NUS) are substantially better than the other two according to both metrics. This ranking is in contrast to the ranking of systems on the CoNLL-14 shared task test against minimally edited references. On these sentences, AMU, which was tuned to M^2, achieves the highest M^2 score with 49.5 and CAMB16, which was the best on the fluency corpus, ranks third with 39.9.

We find that the extent of changes made in the system output is negatively correlated to the qual-

[3]No conclusive recommendation about the best-suited metric for evaluating fluency corrections can be drawn from these results. With only four systems, there is no significant difference between the metric rankings, and even the human rank has no significant difference between three systems.
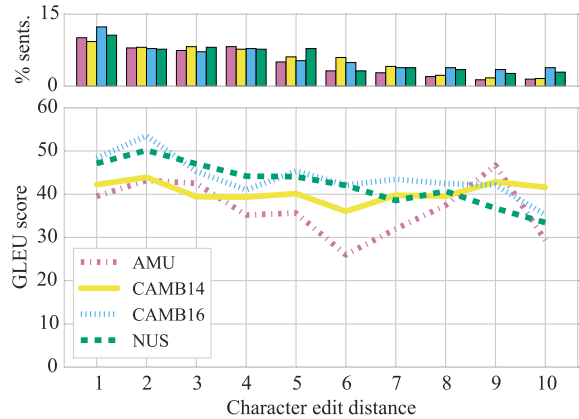

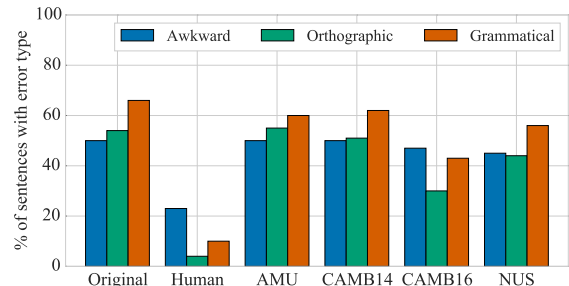Figure 1: GLEU score of system output by LD from input.

Figure 2: Types of errors present in the original, annotated, and system output sentences.

| | | Error type in original | | |
		Awkward	Ortho.	Grammatical
AMU	F	2%	2%	2%
	M	60%	60%	64%
CAMB14	F	2%	0%	2%
	M	64%	69%	65%
CAMB16	F	8%	7%	6%
	M	82%	85%	79%
NUS	F	4%	4%	3%
	M	68%	81%	79%

Table 6: Percent of sentences by error type changed in system output with fluency (F) and minimal (M) edits.

ity as measured by GLEU (Figure 1). The neural systems have the highest scores for nearly all edit distances, and generate the most sentences with higher LDs. CAMB14 has the most consistent GLEU scores. The AMU scores of sentences with LD > 6 are erratic due to the small number of sentences it outputs with that extent of change.

5 Qualitative analysis

We examine the system outputs of the 100 sentences analyzed in Section 3, and label them by the type of errors they contain (Figure 2) and edit types made (Table 6). The system rankings in Table 5 correspond to the rank of systems by the percent of output sentences with errors and the percent of error-ful sentences changed. Humans make significantly more fluency and minimal edits

Original	First , advertissment make me to buy some thing unplanly .
Human	First , an advertisement made me buy something unplanned .
AMU	First , advertissment makes me to buy some thing unplanly .
CAMB14	First , advertisement makes me to buy some things unplanly .
CAMB16	First , please let me buy something bad .
NUS	First , advertissment make me to buy some thing unplanly .
Original	For example , in 2 0 0 6 world cup form Germany , as many conch wanna term work .
Human	For example , in the 2006 World Cup in Germany, many coaches wanted teamwork .
AMU	For example , in the 2 0 0 6 world cup from Germany , as many conch wanna term work .
CAMB14	For example , in 2006 the world cup from Germany , as many conch wanna term work .
CAMB16	For example , in 2006 the world cup from Germany , as many conch , ' work .
NUS	For example , in 2 0 0 6 World Cup from Germany , as many conch wanna term work .

Table 7: Examples of how human and systems corrected GUG sentences.

than any of the systems. The models with neural components, CAMB16 followed by NUS, make the most changes and produce fewer sentences with errors. Systems often change only one or two errors in a sentence but fail to address others. Minimal edits are the primary type of edits made by all systems (AMU and CAMB14 made one fluency correction each, NUS two, and CAMB16 five) while humans use fluency edits to correct nearly 30% of the sentences.

Spelling mistakes are often ignored: AMU corrects very few spelling errors, and even CAMB16, which makes the most corrections, still ignores misspellings in 30% of sentences. Robust spelling correction would make a noticeable difference to output quality. Most systems produce corrections that are meaning preserving, however, CAMB16 changed the meaning of 15 sentences. This is a downside of neural models that should be considered, even though neural MT generates the best output by all other measures.

The examples in Table 7 illustrate some of these successes and shortcomings. The first sentence can be corrected with minimal edits, and both AMU and CAMB14 correct the number agreement but leave the incorrect *unplanly* and the infinitival *to*. In addition, AMU does not correct the spelling of *advertissement* or *some thing*. CAMB16 changes the meaning of the sentence altogether, even though the output is fluent, and NUS makes no changes. The next set of sentences contains many errors and requires inference and fluency rewrites to correct. The human annotator deduces that the last clause is about coaches, not mollusks, and rewrites it grammatically given the context of the rest of the sentence. Systems handle the second clause moderately well but are unable to correct the final clause: only CAMB16 attempts to cor-

rect it, but the result is nonsensical.

6 Conclusions

This paper presents JFLEG, a new corpus for developing and evaluating GEC systems with respect to fluency as well as grammaticality.[4] Our hope is that this corpus will serve as a starting point for advancing GEC beyond minimal error corrections. We described qualitative and quantitative analysis of JFLEG, and benchmarked four leading systems on this data. The relative performance of these systems varies considerably when evaluated on a fluency corpus compared to a minimal-edit corpus, underlining the need for a new dataset for evaluating GEC. Overall, current systems can successfully correct closed-class targets such as number agreement and prepositions errors (with incomplete coverage), but ignore many spelling mistakes and long-range context-dependent errors. Neural methods are better than other systems at making fluency edits, but this may be at the expense of maintaining the meaning of the input. As there is still a long way to go in approaching the performance of a human proofreader, these results and benchmark analyses help identify specific issues that GEC systems can improve in future research.

---

[4]https://github.com/keisks/jfleg

References

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*, pages 697–707, Beijing, China, July. 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, November. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada, June. 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, June. Association for Computational Linguistics.

Mariano Felice, Zheng Yuan, Øistein E. Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 15–24, Baltimore, Maryland, June. Association for Computational Linguistics.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 174–180, Baltimore, Maryland, June. Association for Computational Linguistics.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill^TM: A Bayesian skill rating system. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, pages 569–576, Vancouver, British Columbia, Canada, December. MIT Press.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1546–1556, Austin, Texas, November. 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*, pages 588–593, Beijing, China, July. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria, August. 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, June. Association for Computational Linguistics.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.

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

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, June. Association for Computational Linguistics.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California, June. Association for Computational Linguistics.