Gender Bias and Universal Substitution Adversarial Attacks on Grammatical Error Correction Systems for Automated Assessment

Vygas Raina, Mark Gales
University of Cambridge
{vr313, mjfg}@cam.ac.uk

Extended Abstract

Grammatical Error Correction (GEC) systems perform a sequence-to-sequence task [1], where an input word sequence containing grammatical errors, is corrected for these errors by the GEC system to output a grammatically correct word sequence. With the advent of deep learning methods, automated GEC systems have become increasingly popular. For example, GEC systems are often used on speech transcriptions of English learners as a form of assessment and feedback - these powerful GEC systems can be used to automatically measure an aspect of a candidate’s fluency. The count of edits from a candidate’s input sentence (or essay) to a GEC system’s grammatically corrected output sentence is indicative of a candidate’s language ability, where fewer edits suggest better fluency. The count of edits can thus be viewed as a fluency score with zero implying perfect fluency. However, although deep learning based GEC systems are extremely powerful and accurate, they are susceptible to adversarial attacks: an adversary can introduce a small, specific change at the input of a system that causes a large, undesired change at the output [2]. When considering the application of GEC systems to automated language assessment, the aim of an adversary could be to cheat by making a small change to a grammatically incorrect input sentence that conceals the errors from a GEC system, such that no edits are found and the candidate is unjustly awarded a perfect fluency score.

Nevertheless, most adversarial attack generation approaches in literature require multiple queries of the target system [3, 4]. However, in the setting of language assessment, a candidate cannot query a GEC system. To overcome this issue, this work uses universal adversarial attacks [5], where the same small change has to be made to any input sentence, such that the errors are concealed from the GEC system to obtain a perfect fluency score. As the candidates are non-native speakers of English, it is further required that the form of the attack has to be simple to apply. The simplest such attack is in the form of universal substitutions to exploit potential gender biases in a GEC system. For example, a candidate could replace all male pronouns with female pronouns, e.g. any occurrence of he is replaced with she. To determine the extent of threat of this form of adversarial attack, experiments were performed using a popular, publicly available Transformer-based GEC system, the Gramformer [6], when applied to three benchmark GEC datasets [7, 8, 9], shown in Table 1.

The impact of a universal gender pronoun substitution attack is shown in Table 2. For all datasets the GEC system is worryingly biased by the gender, where a candidate can reduce the number of edits made by the GEC system by simply swapping all male gender pronouns with female pronouns (m2f).

The gender pronoun substitution attack can be generalized to a universal substitution attack: a fixed dictionary mapping of word substitutions can be defined for some target words. When a target word appears in an input sequence it is replaced with its corresponding substitution word. For automated assessment with GEC, an adversary can learn and define the optimal dictionary of word mappings that when applied to any input deceives the GEC system into making no edits. The adversary can sell this dictionary to candidates looking to engage in malpractice - this is a universal substitution attack approach that is agnostic to the original input sequence.

To mimic a realistic setting, the universal substitution dictionary is learnt using only the FCE train set and impact of the adversarial attack is evaluated on other test sets. For computational feasibility, the number of target words has to be limited, as identification of the optimal substitution word demands a greedy search through the English vocabulary. Selection of target words is thus hand-crafted: the most frequent words in the FCE train set, separately for each part of speech (POS), are identified. The universal learnt substituted words are matched in POS with the target words they replace. In this work, target words are restricted to nouns, adjectives or adverbs, e.g. it is found that the target noun life should be substituted with the noun metamorphosis to reduce number of edits. Table 3 presents the impact of the universal substitution dictionary when applied to the unseen BEA and CoNLL test sets, where the dictionary has only a total of 14 target words (6 nouns, 4 adjectives, 2 adverbs and 3 gender pronouns). Note that results are presented only for the samples that are affected by the substitutions. It is interesting to note that even with such few target words there is a reduction in the number of edits made by the GEC system on unseen test sets.

Table 1: GEC system performance

|        | FCE | BEA | CoNLL |
|--------|-----|-----|-------|
| F1 (%) | 49.8| 45.2| 37.1  |

Table 2: Change (%) in Avg. Edits with gender substitution.

|        | FCE | BEA | CoNLL |
|--------|-----|-----|-------|
| $m_m$  | $-7.2\%$ | $-2.8\%$ | $-0.5\%$ $↓$ |
| $f_m$  | $+64.3\%$ | $+15.3\%$ | $+14.8\%$ $↑$ |

Table 3:Avg. number of GEC edits with Universal attack.

|        | No Attack | Sub Attack |
|--------|-----------|------------|
| BEA    | 2.665     | 2.512      |
| CoNLL  | 2.554     | 2.437      |

This work reports on research supported by Cambridge Assessment, University of Cambridge.
1. References

[1] J. Ebrahimi, D. Lowd, and D. Dou, “On adversarial examples for character-level neural machine translation,” CoRR, vol. abs/1806.09030, 2018. [Online]. Available: http://arxiv.org/abs/1806.09030

[2] L. Wang and X. Zheng, “Improving grammatical error correction models with purpose-built adversarial examples,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, Nov. 2020, pp. 2858–2869. [Online]. Available: https://aclanthology.org/2020.emnlp-main.228

[3] W. E. Zhang, Q. Z. Sheng, and A. Alhazmi, “Generating textual adversarial examples for deep learning models: A survey,” CoRR, vol. abs/1901.06796, 2019. [Online]. Available: http://arxiv.org/abs/1901.06796

[4] V. Raina and M. Gales, “Residue-based natural language adversarial attack detection,” in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 3836–3848. [Online]. Available: https://aclanthology.org/2022.naacl-main.281

[5] V. Raina, M. J. F. Gales, and K. Knill, “Universal adversarial attacks on spoken language assessment systems,” in INTERSPEECH, 2020.

[6] P. Damodaran, “Prithivirajdamodaran/gramformer: A framework for detecting, highlighting and correcting grammatical errors on natural language text. created by prithiviraj damodaran. open to pull requests and other forms of collaboration.” [Online]. Available: https://github.com/PrithivirajDamodaran/Gramformer

[7] H. Yannakoudakis, T. Briscoe, and B. Medlock, “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. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 180–189. [Online]. Available: https://aclanthology.org/P11-1019

[8] C. Bryant, M. Felice, Ø. E. Andersen, and T. Briscoe, “The BEA-2019 shared task on grammatical error correction,” in Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 52–75. [Online]. Available: https://aclanthology.org/W19-4406

[9] H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant, “The CoNLL-2014 shared task on grammatical error correction,” in Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 1–14. [Online]. Available: https://aclanthology.org/W14-1701
A. Substitution Search Words

Table A.1 enumerates the target (most frequent) search words from the FCE training set. These words were targeted for the universal substitutions for each part of speech (POS). The FCE training set sentences were used to greedily learn the substitution words for each target word, where the selected word is one that minimises GEC edits over the training set sentences. The FCE test set was used to identify the successful and unsuccessful universal substitutions, as given in colour-coded Tables: A.3, A.4, A.5, A.6 and A.7.
### Table A.1: 8 most common words by POS tag for FCE training set (Filtered to only contain grammatically incorrect sentences). Words with fewer than 100 occurrences are omitted.

| Tag  | Description                  | Words                                                                 |
|------|------------------------------|-----------------------------------------------------------------------|
| CC   | conjunction                 | and, but, or, But, And                                                |
| CD   | numeral cardinal            | one, two, 647 393                                                     |
| DT   | determiner                  | the, a, this, some, The, all, that, an, 12994 5209 1446 1001 966 983 719 660 |
| EX   | existential there           | there, There, 768 174                                                 |
| IN   | preposition                 | in, of, for, that, because, at, with, on, 5045 5106 8200 2940 2054 1896 1705 1532 |
| JJR  | adjective comparative       | more, better, 473 111                                                |
| JJS  | adjective superlative       | best, most, 245 102                                                 |
| JJ   | adjective                  | good, other, last, different, many, great, much, new, 788 573 393 356 355 349 343 312 |
| MD   | modal auxiliary            | would, will, could, should, must, ca, may, 2162 1525 1347 974 553 217 203 102 |
| NNP  | proper noun                | July, London, Pat, Danny, THE, Broo, First, TO                       |
| NNS  | noun plural                | people, clothes, things, activities, years, friends, students, discounts |
| NN   | common noun                | show, time, money, life, school, advertisement, shopping, lot        |
| PDT  | pre-determiner             | all, 447                                                             |
| POS  | genitive marker            | 's, 503                                                             |
| PRP  | pronoun personal           | I, you, it, we, me, they, It, them, 14400 4364 3875 2165 1949 1061 854 615 |
| PRPS | pronoun possessive         | my, your, our, their, her, his, My, Your, 3006 1564 925 448 333 289 169 143 |
| RBR  | adverb comparative         | more, 426                                                           |
| RBS  | adverb superlative         | most, 255                                                            |
| RB   | adverb                      | n't, not, very, so, also, really, only, just, 1852 1774 1609 740 583 554 548 399 |
| RP   | particle                   | up, out, 347 301                                                     |
| VBD  | verb past tense            | was, had, were, did, went, started, said, told, 4627 1358 1026 482 390 378 354 244 |
| VBG  | verb present participle     | going, writing, looking, shopping, being, doing, evening, playing, 496 378 239 154 162 151 128 110 |
| VBN  | verb past participle       | been, closed, seen, changed, written, done, 637 277 143 141 122 103 |
| VBP  | verb present               | have, are, am, do, think, 'm, want, need, 2252 1610 1329 805 709 593 484 327 |
| VB   | verb                       | be, like, have, go, do, know, see, take, 2267 1165 974 702 627 527 463 425 |
| VBJ  | verb 3rd p singular        | is, 's, has, does, 3310 522 428 104                                     |
| WDT  | WH-deteminer               | which, that, 896 382                                                  |
| WP   | WH-pronoun                 | what, who, What, 847 438 127                                         |
| WRB  | WH-adverb                  | when, how, When, where, why, 871 542 281 218 215                     |

### Table A.2: Vocab size for each POS

| Tag  | CC | CD | DT | IN | JJR | JJS | MD | NNS | NN | PRP | PRPS | RBR | RBS | RB | RP | VBD | VBG | VBN | VBP | VB | VB2 | WDT | WP | WRB | WPS |
|------|----|----|----|----|-----|-----|----|-----|----|-----|------|-----|-----|----|----|-----|-----|-----|-----|----|-----|-----|----|-----|-----|
|      | 5  | 12 | 17 | 13 | 39  | 55  | 2300 | 12  | 8088| 29355| 16  | 7   | 4   | 1204 | 278 | 3166| 3022| 2  | 157 | 30  | 3  |
Table A.3: Universal substitution attack on finetuned Gramformer with $N$ most common JJ POS substituted. Results here on FCE test set. Average edits with $N^*$ filter data points that contain the target words substituted.

| N  | Orig | Sub        | ALL        | N1        | N2        | N3        | N4        | N5        | N6        |
|----|------|------------|------------|-----------|-----------|-----------|-----------|-----------|-----------|
| 0  | -    | -          | 1.428 ± 1.752 | 2.069 ± 2.677 | 2.100 ± 2.247 | 2.129 ± 2.122 | 2.112 ± 2.081 | 2.122 ± 2.357 | 1.903 ± 1.850 |
| 1  | good | cavernous  | 1.426 ± 1.746 | 2.300 ± 2.316 | 2.498 ± 2.304 | 2.247 ± 2.235 | 2.672 ± 2.116 | 2.300 ± 2.149 | 2.300 ± 2.316 |
| 2  | other | extraterrestrial | 1.431 ± 1.751 | 2.069 ± 2.081 | 2.129 ± 2.122 | 2.112 ± 2.081 | 2.122 ± 2.357 | 1.903 ± 1.850 | 1.897 ± 2.247 |
| 3  | last | last       | 1.431 ± 1.751 | 2.069 ± 2.081 | 2.129 ± 2.122 | 2.112 ± 2.081 | 2.122 ± 2.357 | 1.903 ± 1.850 | 1.710 ± 2.704 |
| 4  | different | dubious  | 1.429 ± 1.751 | 2.069 ± 2.081 | 2.129 ± 2.122 | 2.112 ± 2.081 | 2.122 ± 2.357 | 1.903 ± 1.850 | 1.897 ± 2.247 |
| 5  | many | dubious    | 1.424 ± 1.751 | 2.069 ± 2.081 | 2.129 ± 2.122 | 2.112 ± 2.081 | 2.122 ± 2.357 | 1.903 ± 1.850 | 1.897 ± 2.247 |
| 6  | great | geopolitical | 1.423 ± 1.748 | 2.069 ± 2.081 | 2.129 ± 2.122 | 2.112 ± 2.081 | 2.122 ± 2.357 | 1.903 ± 1.850 | 1.897 ± 2.247 |
| 7  | much | much       | 1.423 ± 1.748 | 2.069 ± 2.081 | 2.129 ± 2.122 | 2.112 ± 2.081 | 2.122 ± 2.357 | 1.903 ± 1.850 | 1.897 ± 2.247 |

Table A.4: Universal substitution attack on finetuned Gramformer with $N$ most common CC POS substituted. Results here on FCE test set. Average edits with $N^*$ filter data points that contain the target words substituted.

| N  | Orig | Sub        | Edits ALL | Edits ALL |
|----|------|------------|-----------|-----------|
|    |      |            | #2734     | #2734     |
| 0  | -    | -          | 1.428 ± 1.752 | 2.300 ± 2.316 |
| 1  | and  | and        | 1.428 ± 1.752 | 2.300 ± 2.316 |
| 2  | but  | but        | 1.428 ± 1.752 | 2.300 ± 2.316 |
| 3  | or   | or         | 1.428 ± 1.752 | 2.300 ± 2.316 |

Table A.5: Universal substitution attack on finetuned Gramformer with $N$ most common NN POS substituted. Results here on FCE test set. Average edits with $N^*$ filter data points that contain the target words substituted.

| N  | Orig | Sub        | Edits ALL | Edits N6 |
|----|------|------------|-----------|----------|
|    |      |            | #2734     | #9       |
| 0  | -    | -          | 1.428 ± 1.752 | 2.778 ± 3.420 |
| 1  | my   | my         | 1.428 ± 1.752 | 2.778 ± 3.420 |
| 2  | your | your       | 1.428 ± 1.752 | 2.778 ± 3.420 |
| 3  | our  | our        | 1.428 ± 1.752 | 2.778 ± 3.420 |
| 4  | their | their      | 1.428 ± 1.752 | 2.778 ± 3.420 |
| 5  | her  | her        | 1.428 ± 1.752 | 2.778 ± 3.420 |
| 6  | his  | my         | 1.427 ± 1.754 | 2.667 ± 3.391 |

Table A.6: Universal substitution attack on finetuned Gramformer with $N$ most common PRPS POS substituted. Results here on FCE test set. Average edits with $N^*$ filter data points that contain the target words substituted.
Table A.7: Universal substitution attack on finetuned Gramformer with $N$ most common RB POS substituted. Results here on FCE test set. Average edits with $N^*$ filter data points that contain the target words substituted.

| N | Orig | Sub | Edits ALL | Edits N1 | Edits N3 | Edits N4 | Edits N6 |
|---|------|-----|-----------|---------|---------|---------|---------|
|   | -    | -   | #2734 samples | #181 samples | #75 samples | #28 samples | #69 samples |
| 0 | -    | -   | 1.428±1.752 | 2.061±2.317 | 2.027±1.708 | 2.107±1.618 | 1.783±1.688 |
| 1 | very | stylistically | 1.428±1.761 | 2.122±2.401 | 1.947±1.734 | 2.030±1.754 | 1.797±1.820 |
| 2 | so   | so   | 1.430±1.761 | 1.430±1.760 | 1.430±1.760 | 1.430±1.760 | 1.430±1.760 |
| 3 | also | noticeably | 1.430±1.761 | 1.430±1.760 | 1.430±1.760 | 1.430±1.760 | 1.430±1.760 |
| 4 | really | romantically | 1.430±1.761 | 1.430±1.760 | 1.430±1.760 | 1.430±1.760 | 1.430±1.760 |
| 5 | only | only | 1.430±1.761 | 1.430±1.760 | 1.430±1.760 | 1.430±1.760 | 1.430±1.760 |
| 6 | just | passionately | 1.430±1.761 | 1.430±1.760 | 1.430±1.760 | 1.430±1.760 | 1.430±1.760 |

Table A.8: FCE: Universal gender substitution attack