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

UNIVERSITY OF CAMBRIDGE

V. Raina, M.J.F. Gales {vr313,mjfg}@eng.cam.ac.uk

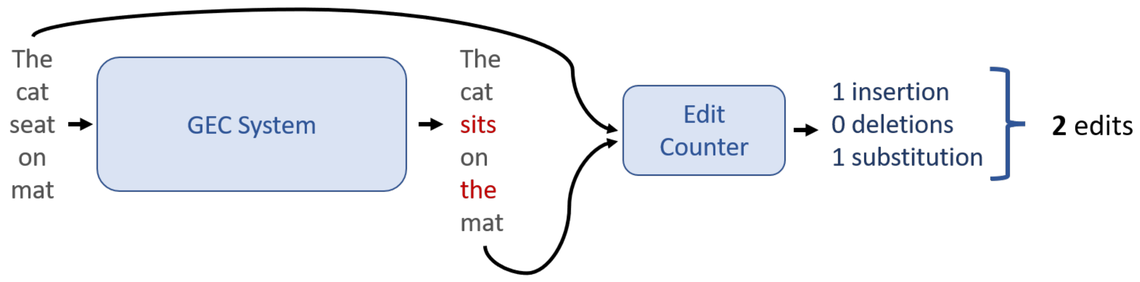
ALTA Institute / Department of Engineering, University of Cambridge

1. Introduction

- Increased demand for English learning and assessment.
- Automated Grammatical Error Correction (GEC) systems can be used for assessment, e.g. GEC on audio transcripts.
- Candidates can engage in mal-practice by adversarial attacks of GEC systems.

2. Grammatical Error Correction for Assessment

- GEC systems perform a sequence-to-sequence task.
  $\hat{y}_{1:L} = \arg\max_{y_{1:L}} \{p(y_{1:L}|x_{1:T}; \theta)\}$

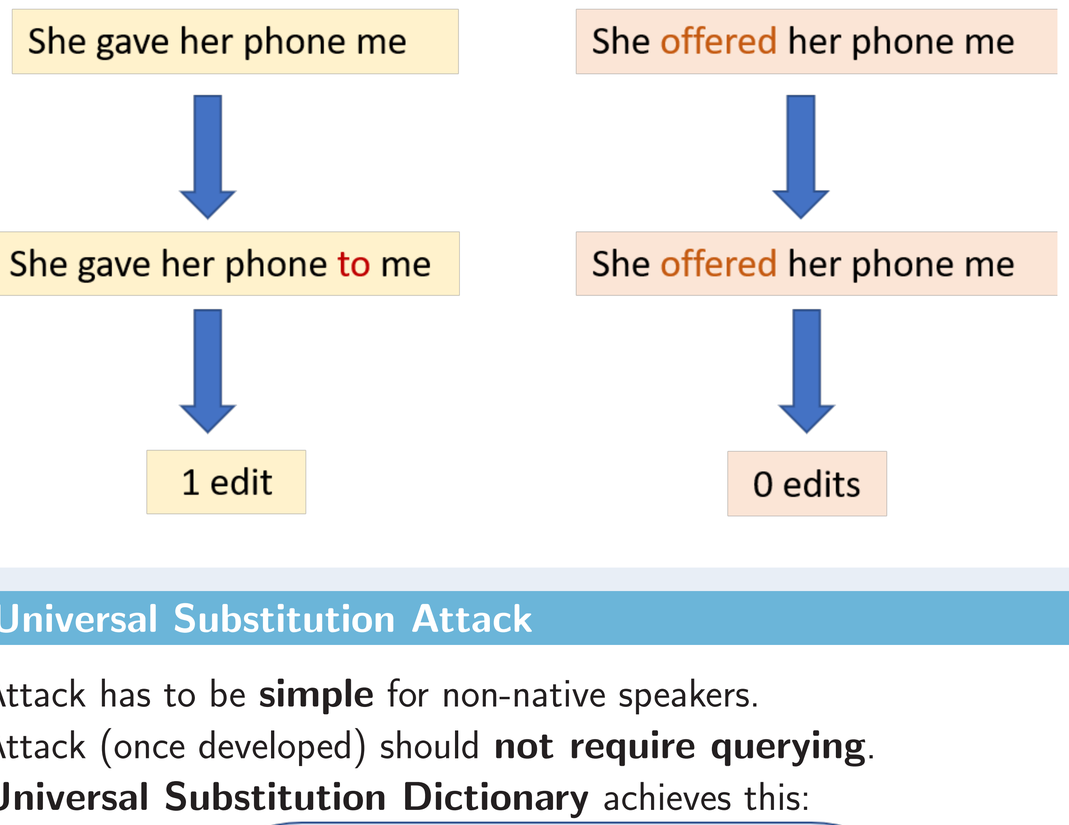


- Number of **edits** is indicative of a candidate's fluency score.
  $\hat{e}_{1:P} = \texttt{edits}(x_{1:T}, \hat{y}_{1:L}) \qquad S_\theta(x_{1:T}) = \texttt{count}(\hat{e}_{1:P}) = P$

3. Adversarial Attack GEC System

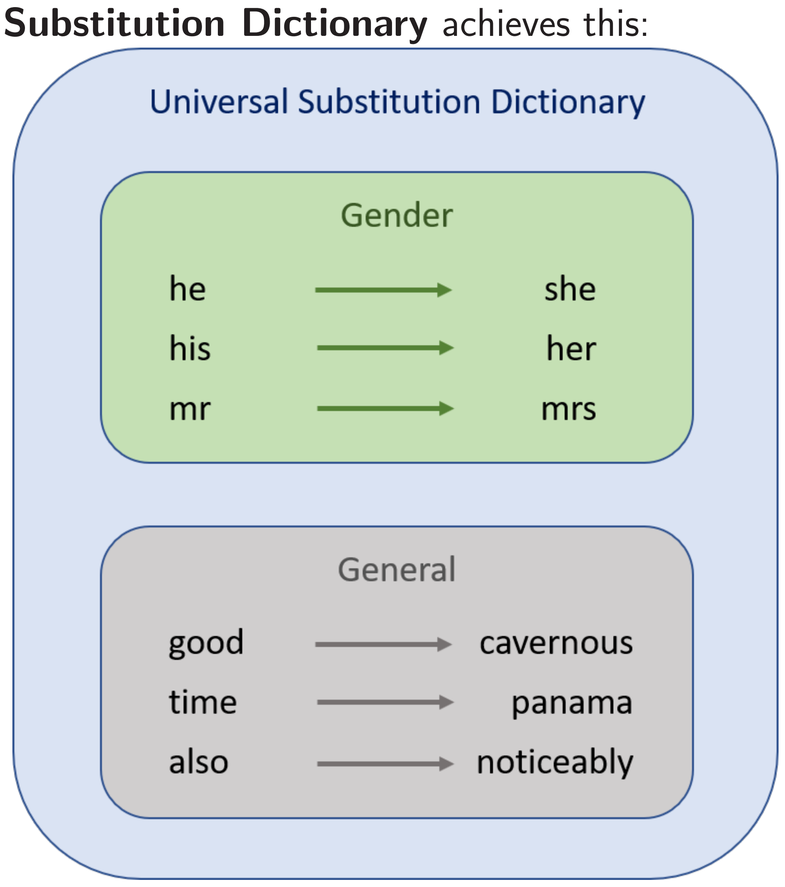
- Adjust input to **deceive** the GEC system into making no edits $\Longrightarrow$ **perfect** fluency score.

$$S_\theta(x'_{1:T'}) = 0 < S_\theta(x_{1:T}) \quad \text{s.t. } \mathcal{H}(x_{1:T}, x'_{1:T'}) \leq \epsilon.$$



4. Universal Substitution Attack

- Attack has to be **simple** for non-native speakers.
- Attack (once developed) should **not require querying**.
- **Universal Substitution Dictionary** achieves this:



- e.g. *He had a good time.* $\rightarrow$ *She had a cavernous panama.*

5. Experiments

- **Gramformer**: online GEC model used for Grammatical Error Correction.
- It is based on the encoder-decoder T5 Transformer.

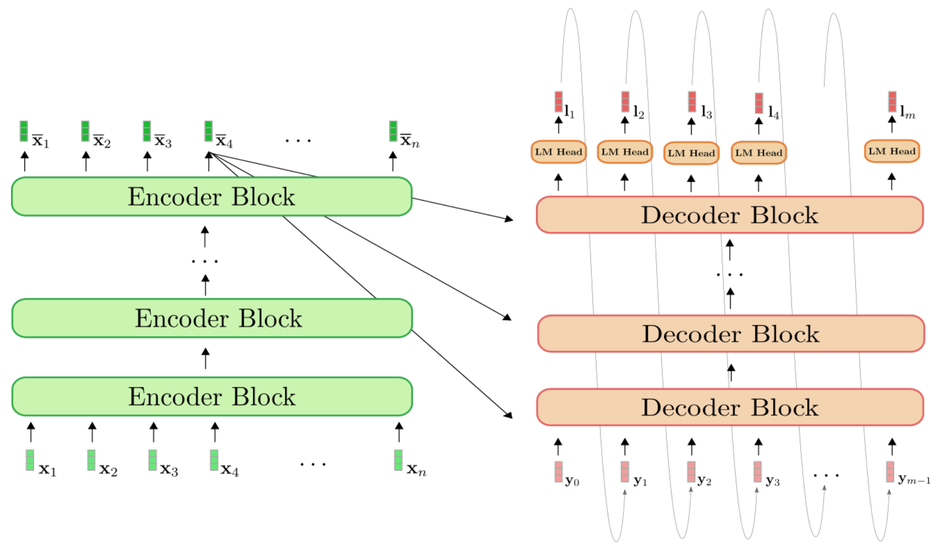


Figure: Source: huggingface[a]

- 3 standard GEC benchmark datasets used for evaluation.

	Precision	Recall	F0.5
FCE	51.6	43.7	49.8
CoNLL-14	49.3	34.1	45.2
BEA-19	35.3	44.6	37.1

6. Results

- The GEC system is **biased** to the traditionally female gender pronouns, i.e., hypothesizes fewer edits.
- % change in **average edits** with gender substitutions, male-to-female (`m2f`) and female-to-male (`f2m`) given below.

	Substitution	FCE	BEA-19	CoNLL-14
−	m2f	-7.2%	-2.8%	-0.5%
+	f2m	+64.3%	+15.3%	+14.8%

- FCE train data used to learn a general Universal Substitution Dictionary (**USD**).
- Most **frequent words** (for each part of speech) selected as **target** words for USD.
- USD evaluated on FCE test data and only **successful** substitutions kept.
- Impact of this evaluated on CoNLL-14 and BEA-19, as measured by the average number of GEC edits from the input to output sequence.
- USD had substitutions for 6 nouns, 4 adjectives, 2 adverbs and 3 gender pronouns.

	Data	Original	Attack
−	CoNLL-14	2.554	2.437
−	BEA-19	2.665	2.512

7. Conclusions

- Automated Grammatical Error Correction Systems play a useful role in language learning and **assessment**.
- State of the art **deployed** GEC systems (in high stakes environments) are susceptible to simple forms of mal-practice. Candidates can cheat by making use of simple **gender biases** and **universal substitution dictionaries** to deceive GEC systems into making **no corrections**, artificially suggesting **perfect fluency**.
- These universal attacks are **agnostic** to the specific input across **multiple datasets**.
- Future work will explore methods to defend GEC systems to ensure **robustness** to adversarial attacks.

[a] https://github.com/huggingface/blog/blob/main/warm-starting-encoder-decoder.md