How Good (really) are Grammatical Error Correction Systems?

Alla Rozovskaya
Queens College, CUNY

Dan Roth
University of Pennsylvania

Problem with Reference-Based Evaluation

- The set of possible golds (space of valid corrections) for a given source sentence is extremely large
- Most GEC datasets contain 1 gold for a given source sentence
  - This (**random**) gold is generated relative to the source sentence
  - The gold is independent of the system output
- **Impact**
  - **Evaluation:** reference-based evaluation underestimates system performance
  - **Training** is also affected as it is performed relative to a single reference

We propose the notion of **Closest Gold**, and study the implications of evaluating relative to it.

Standard Reference-Based Evaluation with Reference Gold (RG)

Source	The settings are very **reallistic** and the actors **had a great** performance .
Reference Gold (RG)	The settings are very realistic and the actors *gave* a great performance .
Hypothesis 1	The settings are very *realistic* and the actors *had great* performance.

Gold edits: (1) reallistic -> realistic;
(2) had -> gave

System edits: (1) reallistic -> realistic;
(2) had a great -> had great

Correct edits: (1) reallistic -> realistic

Precision: 1/2=0.5
Recall: 1/2=0.5

Evaluation with Closest Golds

- **Closest Golds (CGs)** are generated relative to system hypotheses
  - Annotators generate correct text that is closest to the system output
  - We generate CGs for top hypothesis and hypotheses at lower ranks
- CGs are used to evaluate system outputs on **4 GEC datasets**
  - 2 English and 2 Russian datasets
- **Major differences in performance** when using CGs instead of RGs
- We claim that **evaluation relative to CGs gives true system performance**

Reference Gold (RG) vs. Closest Gold (CG) in Evaluation

Source	The settings are very **reallistic** and the actors **had a great performance** .
Hypothesis 1	The settings are very *realistic* and the actors *had great* performance .
Reference Gold (RG)	The settings are very realistic and the actors *gave* a great performance .
Closest Gold (CG) to Hypothesis 1	The settings are very realistic and the actors had great *performances* .

Reference Gold:
Gold edits: (1) reallistic -> realistic;
(2) had -> gave

System edits: (1) reallistic -> realistic;
(2) had a great -> had great

Correct edits: (1) reallistic -> realistic

Precision: 1/2=0.5
Recall: 1/2=0.5

Closest Gold:
Gold edits: (1) reallistic -> realistic;
(2) had a great -> had a great
(3) performance -> performances

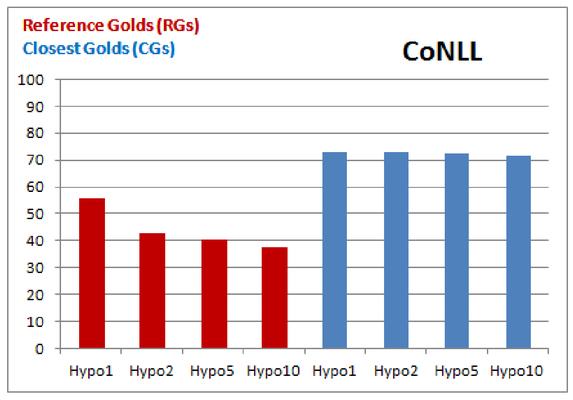
System edits: (1) reallistic -> realistic;
(2) had a great -> had great

Correct edits: (1) reallistic -> realistic
(2) had a great -> had great

Precision: 2/2=1.0
Recall: 2/3=0.66

Key Results

- System performance when evaluated relative to Reference Golds (RGs) is severely underestimated
- Lower rank hypotheses are often as good as the top hypothesis (relative to their CGs)
  - And are more "interesting"



- Evaluation against RGs shows **a large gap between top hypothesis and lower-ranked hypotheses**.
- Evaluation against CGs reveals **very little degradation** between top hypothesis and the rest

Lower-Ranked Hypotheses Propose More Changes

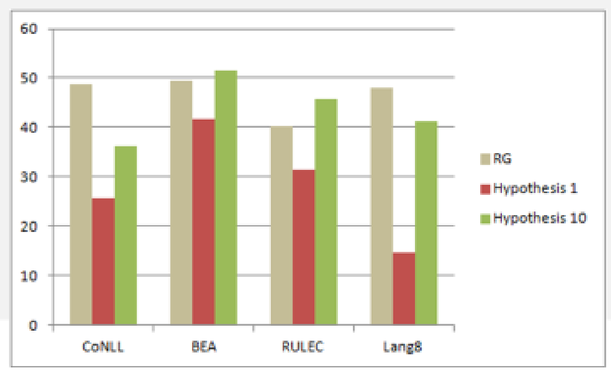
Hypothesis	RULEC (Ru)	Lang8 (Ru)	BEA (En)	CoNLL (En)
H_1	90	98	125	156
H_2	144	186	180	203
H_5	174	214	200	239
H_{10}	194	225	220	266
RG	202	232	202	289

*Number of **edits proposed by the system** (by hypothesis rank). Last row shows number of gold edits in the reference gold.*

- Under-correction phenomenon:
  - The top-ranked hypothesis makes a fraction of edits compared to RGs.
- Lower-ranked hypotheses propose a similar number of changes to RGs

Lower-Ranked Hypotheses Propose More Lexical Changes

- **Top-ranked hypothesis severely under-corrects** compared to humans, especially on lexical errors
- **Lower-ranked hypotheses propose more lexical changes** than top-ranked hypothesis



Percentage of lexical edits relative to the total number of changes.

Conclusion

- Evaluation with *closest golds* has taught us two lessons
  - GEC systems are doing better than standard evaluations show
  - Lower-ranked are interesting and are not better than the top hypothesis
- We propose several recommendations based on these findings (please check out the paper)
  - Evaluation
  - Training and tuning

This work was supported by Contract FA8750-19-2-0201 with the US Defense Advanced Research Projects Agency (DARPA).