



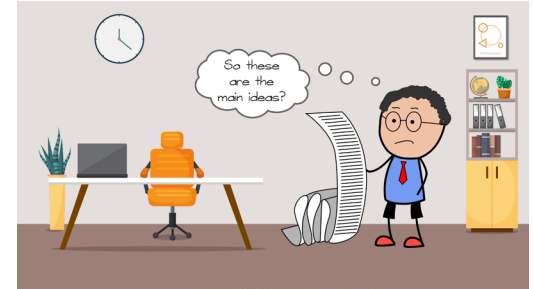
How Good (really) are Grammatical Error Correction Systems?

Alla Rozovskaya and Dan Roth
EACL-2021

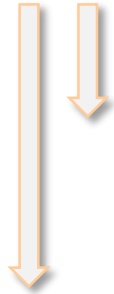


The Multiple Gold Problem

- Many problems do not have a single gold
 - Summarization is an archetypical example
 - A given document has multiple possible summaries
 - Text Correction
 - I gave him a books
- Impacts both the evaluation and the training stage



Standard Reference-Based Evaluation for GEC



Source	The settings are very reallistic and the actors had a great performance .
Reference Gold (RG)	The settings are very <u>realistic</u> and the actors <u>gave</u> a great performance .
Hypothesis 1	The settings are very <u>realistic</u> and the actors <u>had great</u> 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: $1/2=0.5$

Problem with Reference-Based Evaluation

- The set of possible golds (space of valid corrections) for a given source sentence is extremely large (Bryant and Ng, 2015, Choshen and Abend, 2018)
- Most GEC datasets contain 1 (or 2) golds for a source sentence
 - This (**random**) gold is generated relative to the source sentence
 - The reference gold is independent of the system output
- Impact:
 - **Evaluation:** Reference-based evaluation underestimates system performance
 - Also impacts **training** (which is done relative to a single reference gold)


Evaluation with Closest Golds

- Closest Golds (CGs) are generated relative to system hypotheses
 - Annotators generate a correct text that is the closest to the system output
 - CGs are generated for top hypothesis and hypotheses of lower ranks (2, 5, and 10)
- We use closest golds to evaluate system output of 4 GEC datasets
 - 2 English and 2 Russian
- We show **major differences** in performance when using CGs instead of RGs
- We claim that evaluation relative to CGs gives true system performance

Our Key Results

- The system performance, when evaluated relative to reference gold, is severely underestimated
 - And we show by how much
- Lower rank hypotheses are often as good as the top hypothesis (relative to their CGs)
 - And are more “interesting”

Reference-Based Evaluation with Closest Gold



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

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

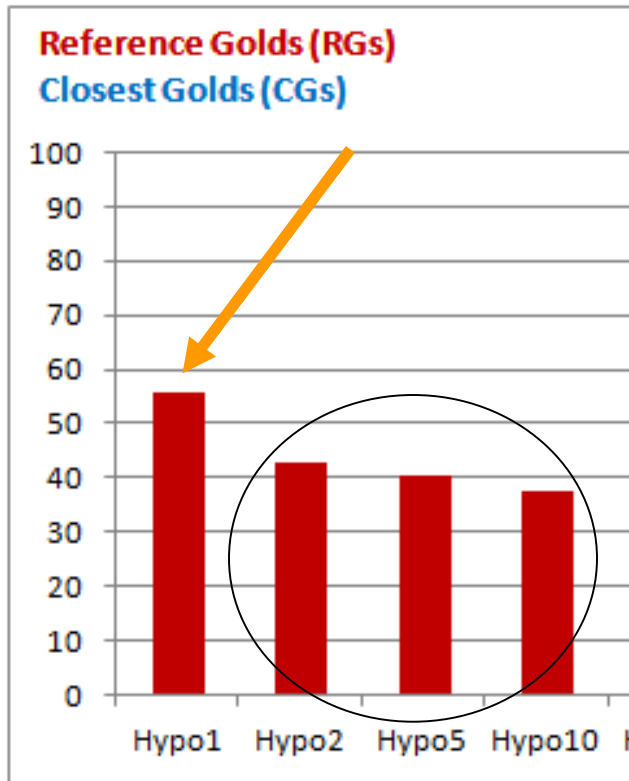
System edits: (1) reallistic -> realistic;
(2) had a great -> had great

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

Precision: $2/2=1.0$

Recall: $2/3=0.66$

Key Findings



More results on other datasets are in the paper

- 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
 - The reason is that lower-ranked hypotheses propose **more diverse changes (e.g. lexical changes)**, that have a **lower chance of matching RGs**

Lower-Ranked Hypotheses Propose More Changes

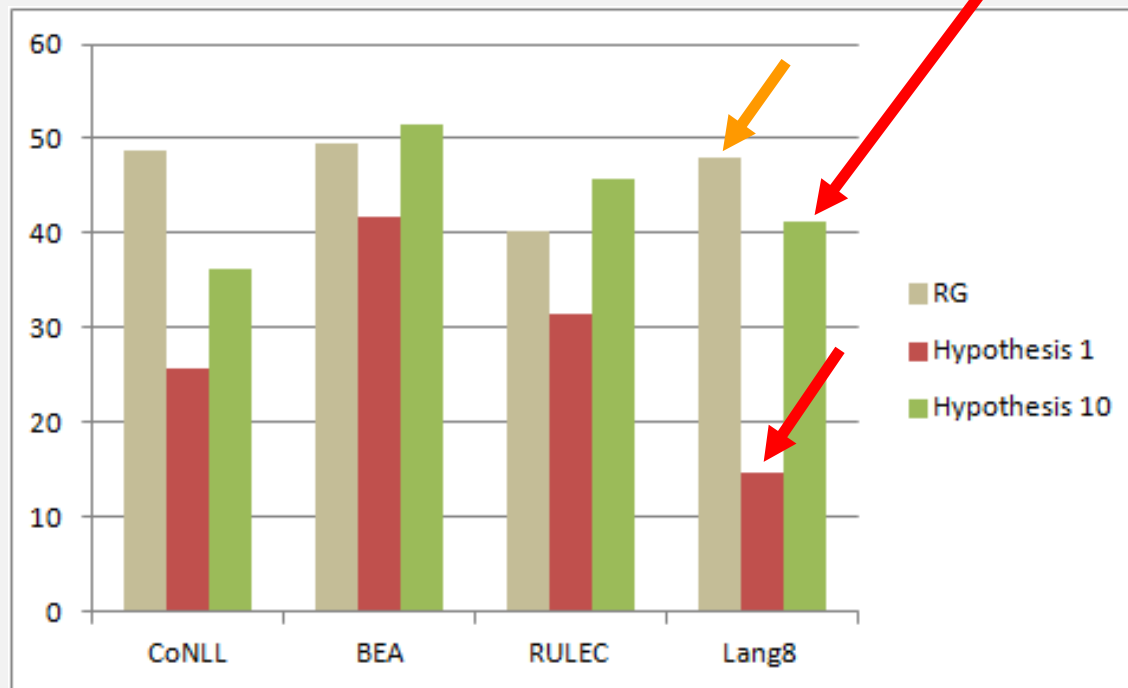
Hypothesis	RULEC (Ru)	Lang8 (Ru)	BEA (En)	CoNLL (En)
H_1	90	98	125	156
H_2	144	186	180	203
H_5	174	214	200	239
H_{10}	194	225	220	266
RG	202	232	202	289

*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 CGs has taught us two lessons
 - We are actually doing better than we thought
 - Lower-ranked hypotheses are interesting and not worse than the top hypothesis
- We propose several recommendations based on these findings (please check out the paper)
 - Evaluation
 - Training and tuning

Thank you!