1000+ valid corrections → Humans score poorly → Under-correction is rewarded → SoTA systems surpass humans

Under-correction is rewarded ← SoTA underpredicts

1000+ valid corrections → More references don't help

SoTA systems surpass humans → SoTA underpredicts

Inherent Biases in Reference-based Evaluation for Grammatical Error Correction and Text Simplification

Leshem Choshen & Omri Abend

School of Computer Science and Engineering & Department of Cognitive Sciences

The Hebrew University of Jerusalem Israel

Reference Based Measures

Number of Valid Corrections

Estimated with crowdsourcing and UnseenEst

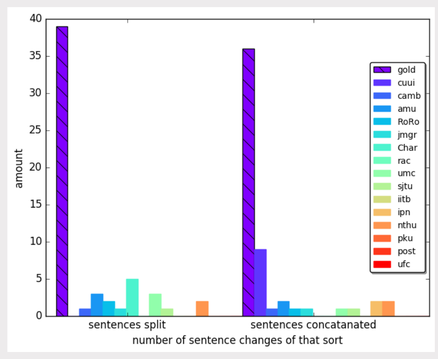
	Frequency Threshold (γ)			
	0	0.001	0.01	0.1
Variants	1351.24	74.34	8.72	1.35
Mass	1	0.75	0.58	0.37

GEC Performance

Under-Prediction (Conservatism)

SoTA systems correct an order of magnitude less than humans

In terms of: word changes, sentence splits/merges and word reordering

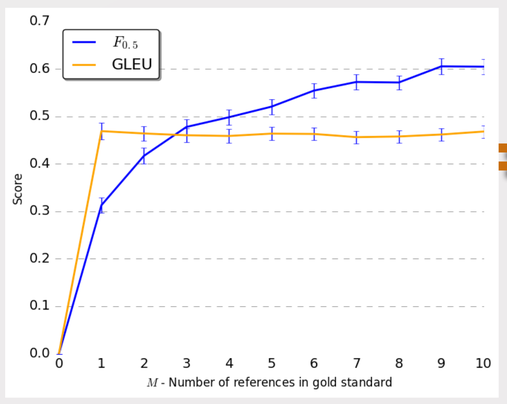


Perfect Correctors (Humans)

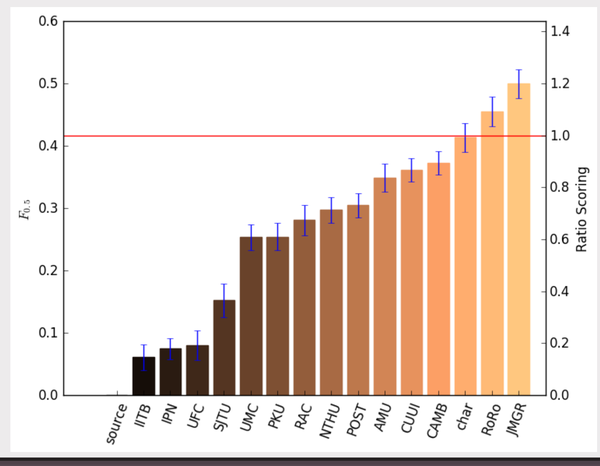
Accuracy, GLEU and M^2

Loss and evaluation metrics assign low scores to perfect correctors

Increasing references won't solve it



Systems on par with Humans



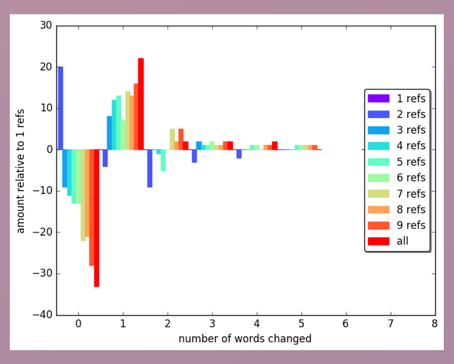
RBMs Favor Some (Valid) Corrections

Coverage ↛Conservatism

And SoTA favors smilar ones

Coverage ↛Conservatism

$P_{correct} \cdot P_{covered} < 1 - P_{detect}$



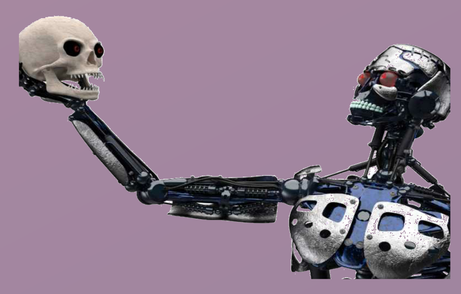
Encourage close-class errors

Discourage open-class errors

Disincentivized to correct-

Even if you know the answer

Precision oriented measures make it worse



What can we do?

Reference-less measures

Beyond n-gram overlap of source\reference (Semantics)

USim [Choshen & Abend 2018, a]

More in the paper

Significance, methodological contributions, Empirical number of corrections per error type [Choshen & Abend 2018, b]

References

a. Choshen & Abend (NAACL 2018)
Reference-less Measure of Faithfulness for Grammatical Error Correction.

b. Choshen & Abend (ACL 2018)
Automatic Metric Validation for Grammatical Error Correction

UCCA Parsing Shared Task - SemEval 2019

WE WANT YOU!