Too many corrections: Semantic reference-less evaluation for Grammatical Error Correction

Leshem Choshen & Omri Abend

Hebrew University Jerusalem Israel

September 25 2017

Overview

Plan

the task

- Input: a text which is perhaps ungramatical
- Output: a grammatical text saying the same meaning/content.

Example: However , there are both sides of stories

The task

- Input: a text which is perhaps ~~ungramatical~~ ungrammatical
- Output: a grammatical text ~~saying~~ conveying the same meaning/content.

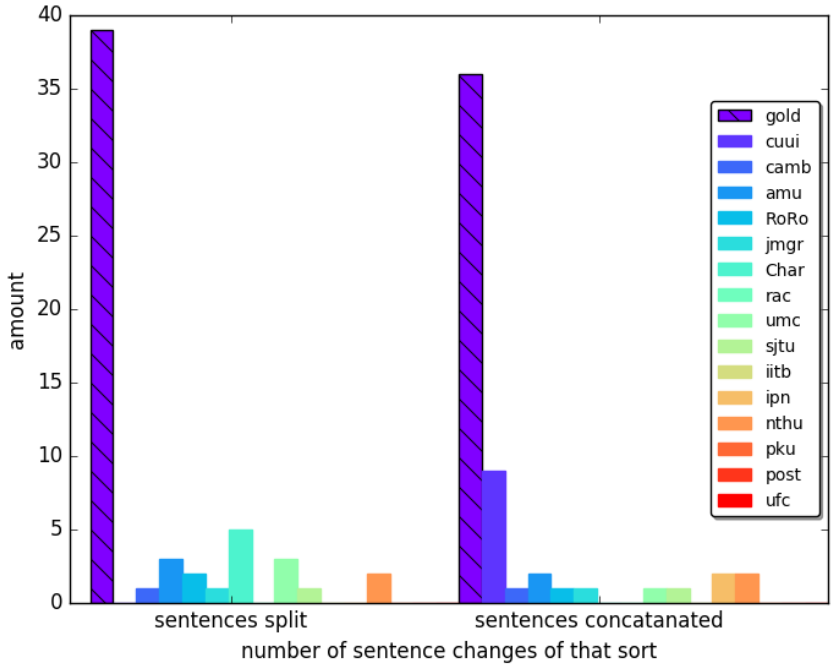
Example: However , there are ~~both sides of stories~~ →
However , there are two sides to every story.

Plan

Current systems hardly change

It is a virtue to avoid bad corrections, but correcting is still the goal...

- Less words changed
- Less word order changes
- Less sentences split into two
- Less sentences merged into one
- and so on...



Plan

What exists

Evaluation measures all share in common:

- Compare a system correction to a set of references.
- Emphasize precision over recall.

Corpora:

- Train and validation - 1 reference per source sentence, never more.

Corrections as distribution

- Each sentence x has a set of valid corrections $correct_x$
- $\mathcal{D}_x$ a distribution of human corrections
- In a Corpus - $Y \sim \mathcal{D}_x^M$ a sample of M references
- $P_{coverage}$ - $P_{y \sim \mathcal{D}_x}(y \in Y)$

Distributions as we get from crowdsourcing

- Hypothesis: Probably more than 2 references, and they are not uniform but approximately so.

Distributions as we get from crowdsourcing

- Hypothesis: Probably more than 2 references, and they are not uniform but approximately so.
- Result: On average 1351.24 corrections per sentence with 8-15 words with a heavy tail like behaviour.

Analytical worries

Oracle chooses whether to produce a correction or not.
Mistake detected: incentivized to correct it only if

$$p_{correct} \cdot p_{coverage} > 1 - p_{detect}$$

Or, given α punishment for wrong corrections

$$p_{correct} \cdot p_{coverage} - (1 - p_{correct} \cdot p_{coverage})\,\alpha > 1 - p_{detect}$$

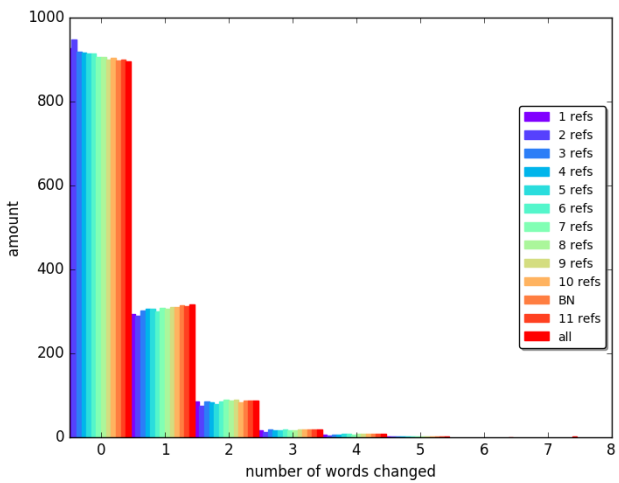
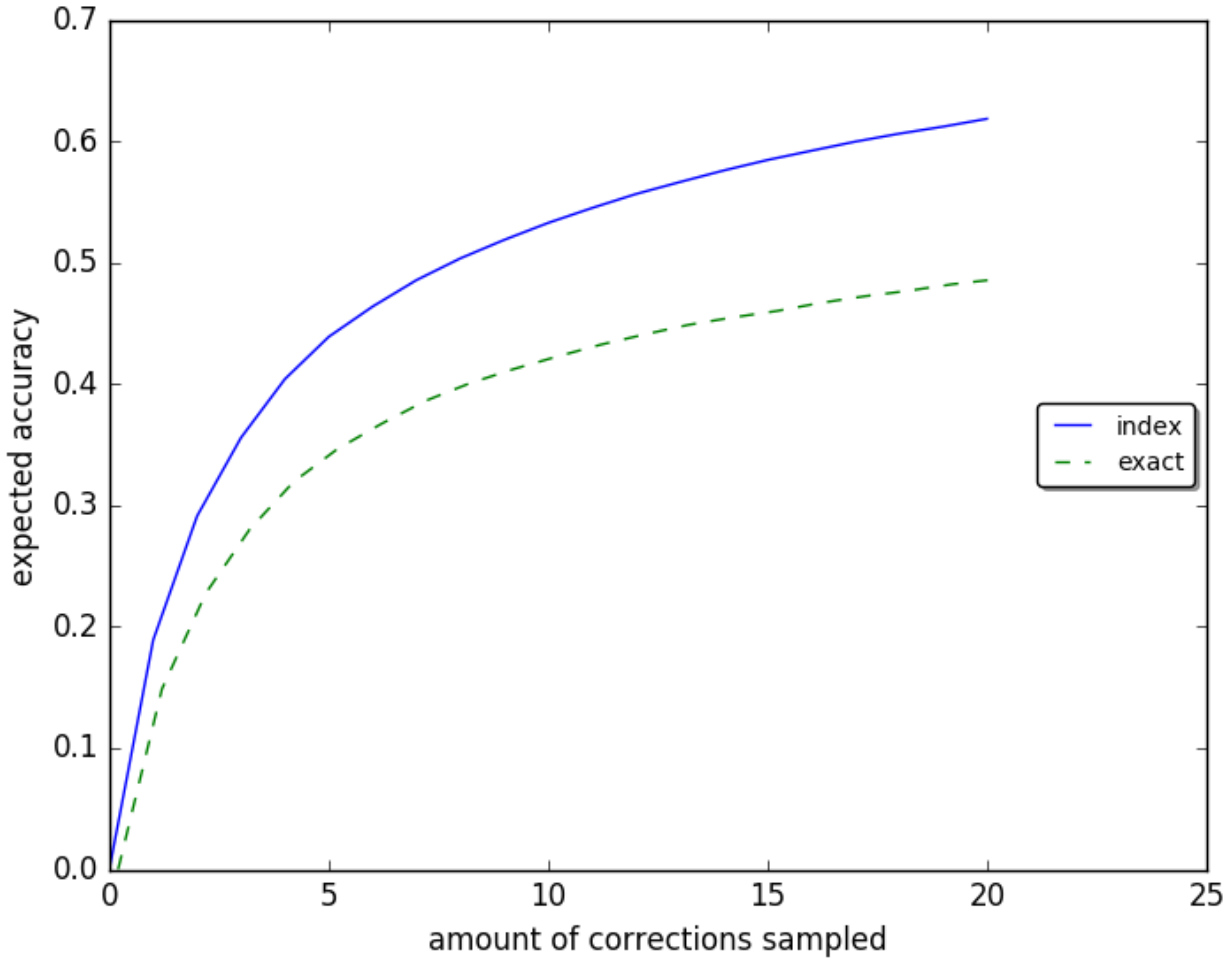


Figure: Also, empirical worries, for decoration

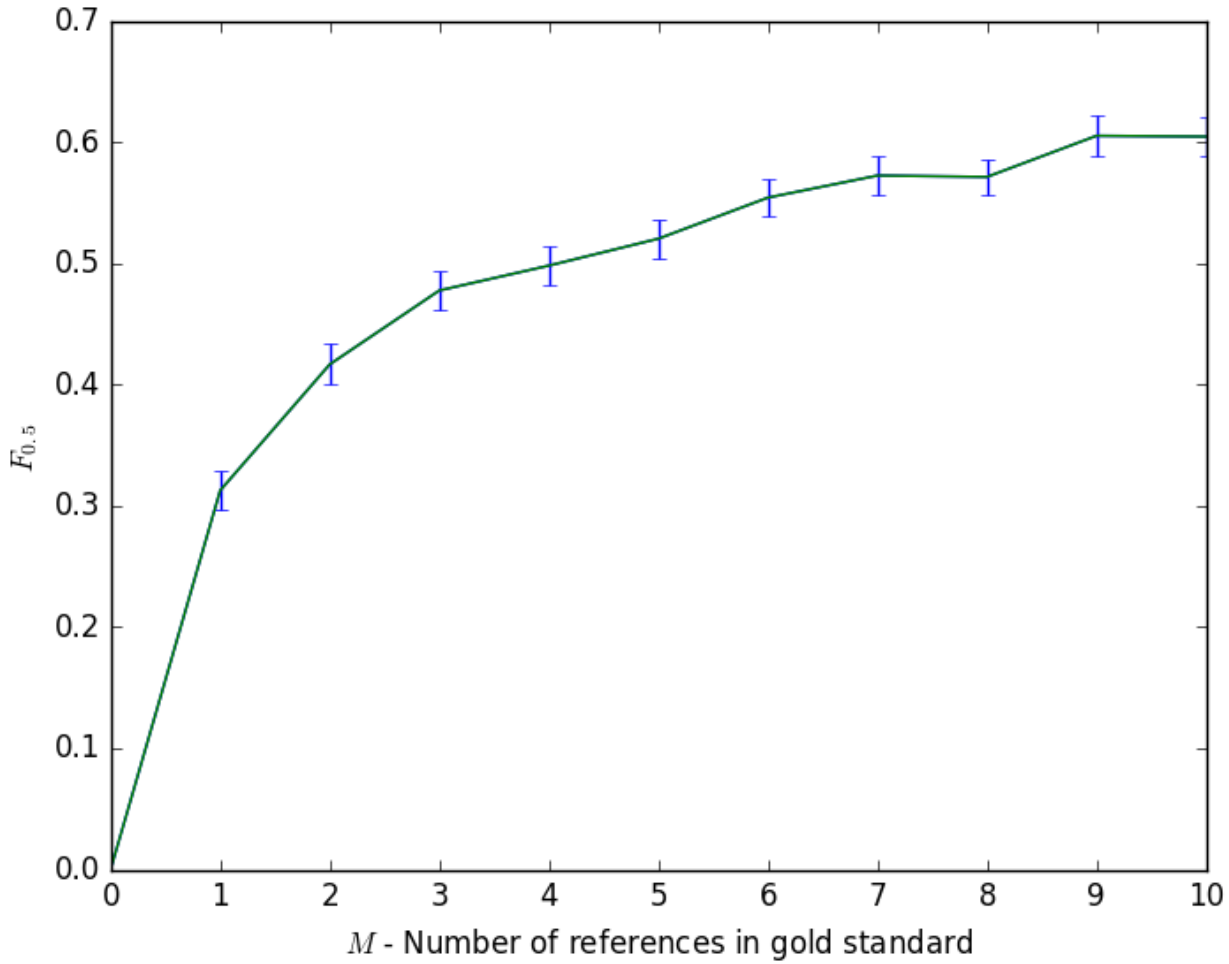
Accuracy - analysis

Given a perfect corrector, how well will it do?

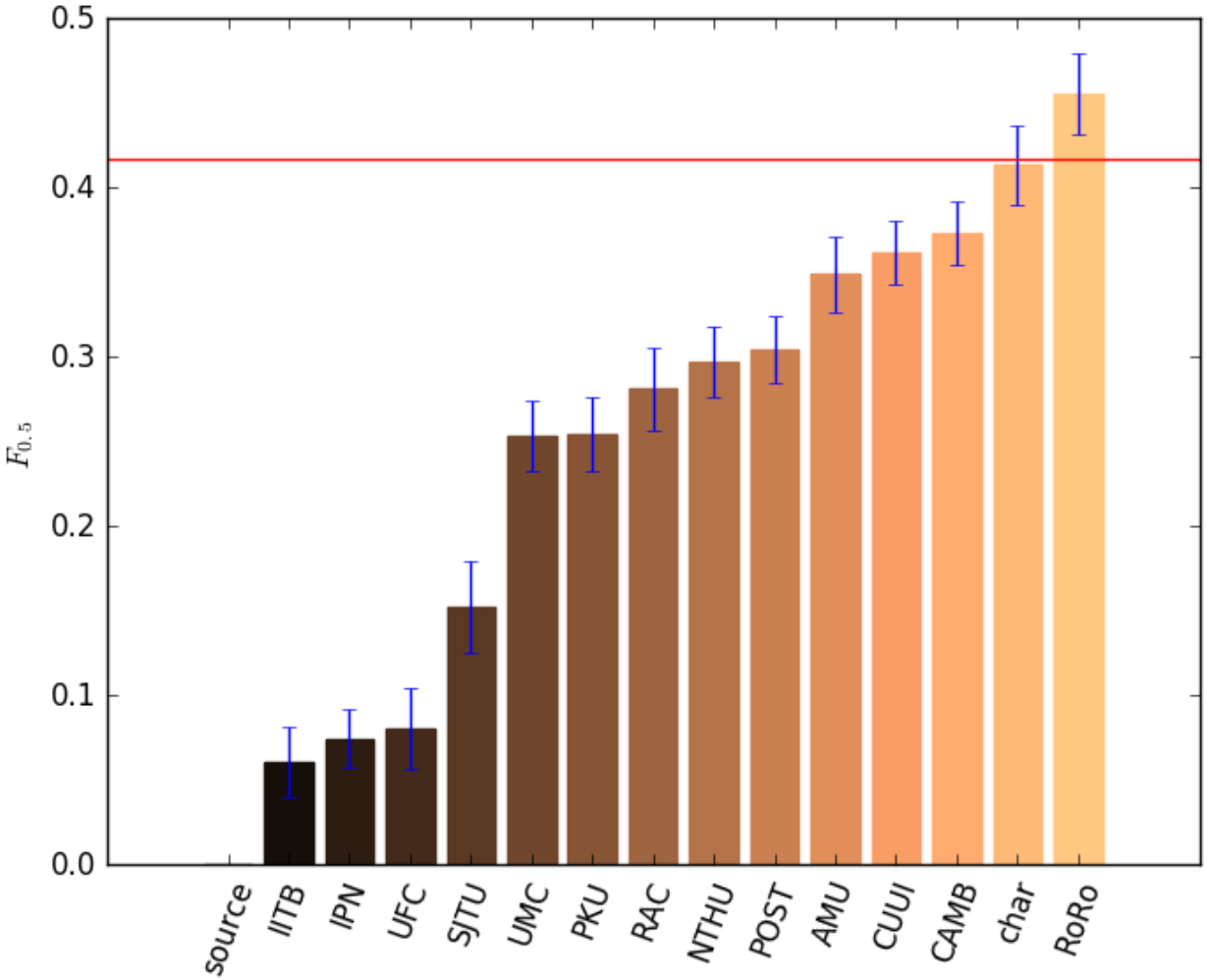
$\frac{1}{N}\sum_{i=1}^{N} P_{Y\sim\mathcal{D}_i^M, y\sim\mathcal{D}_i}(y \in Y)$



$F_{0.5}$-score - empirical



human vs. machine - test on 2 references



Plan

Reference-less evaluation

Input: Corrected sentences and Source sentences ~~and references in the form of sentences~~.
Output: A score, but which?!

Reference-less evaluation

Combine two measures

1. faithfulness – semantic similarity of the correction and the source. [1]
2. grammaticality – error detection over the source [2]

[1] Leshem Choshen and Omri Abend. "Conservatism and Over-conservatism in Grammatical Error Correction" - under revision

[2] Napoles Courtney, Keisuke Sakaguchi, and Joel Tetreault. "There's No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction." arXiv preprint arXiv:1610.02124 (2016).

UCCA

- Semantic annotation scheme that builds on typological and cognitive linguistic theories
- Provides a coarse-grained, cross-linguistically applicable representation
- Structures are DAGS, words are leaves

Ungrammatical hypotheses

- Ungrammatical text can be annotated using UCCA
- Corrections change grammar, not semantics

UCCASim(ilarity) between source and reference

	UCCASim
Different annotators	0.84
Same annotator	0.92
TUPA parser	0.7
Ungrammatical IAA	0.83
Baseline IAA [1]	0.79
TUPA reported precision	0.69

[1] UCCA IAA improved since the original paper

Thank you