Reference-based Metrics can be Replaced with Reference-less Metrics in Evaluating Grammatical Error Correction Systems

Hiroki Asano[1,2], Tomoya Mizumoto[2], Kentaro Inui[1,2]
[1]Graduate School of Information Sciences, Tohoku University
[2]RIKEN Center for Advanced Intelligence Project

Overview

Background

- In grammatical error correction, automatically evaluating GEC systems requires gold-standard references, which tend to be expensive and limited in coverage.
- To address this problem, a reference-less approach has recently emerged [Napoles et al.,2016].
- The metrics, which only consider the criterion of grammaticality, have not worked as well as reference-based metrics.

Summary

- We propose **a reference-less metric** that combines ***fluency*** and ***meaning preservation*** with ***grammaticality***.
- The proposed reference-less metric provides **a better estimate of manual scores** than that of commonly used reference-based metrics.

Source (written by learners of English)
Machine is *design* to help people .

Hypothesis (GEC system's output)
The machine is designed to help people .

~~Reference (Gold-standard)~~
~~Machines are designed to help people .~~

Reference-less GEC assessment

We combined three criteria:

Grammaticality

- For a hypothesis h, grammaticality score $S_G(h)$ determined by a logistic regression with linguistic features:
  - the number of misspellings
  - n-gram language model score
  - PCFG and rink grammar features
  - the number of errors detected by Language Tool

Fluency

- The importance of fluency in GEC has been shown by Sakaguchi et al. (2016).
- Fluency can be captured by statistical language modeling [Lau et al., 2015].
- For a hypothesis h, fluency score $S_F(h)$ is calculated as follows:

$$S_F(h) = \frac{\log P_m(h) - \log P_u(h)}{|h|}$$

P_m: the probability of the sentence given by RNNLM
P_u: the unigram probability of the sentence

Meaning Preservation

- In GEC, the meaning of original sentences should be preserved.
- Without this criterion, a gaming system would not be penalized.
- We adopt METEOR (Denkowski and Lavie, 2014).
- Meaning score $S_M(h,s)$ for a source sentence s and a hypothesis h is calculated as follows.

$$S_M(h,s) = \frac{P \cdot R}{t \cdot P + (1-t) \cdot R},$$

$$where\ P = \frac{m(h_c, s_c)}{|h_c|}, R = \frac{m(h_c, s_c)}{|s_c|}$$

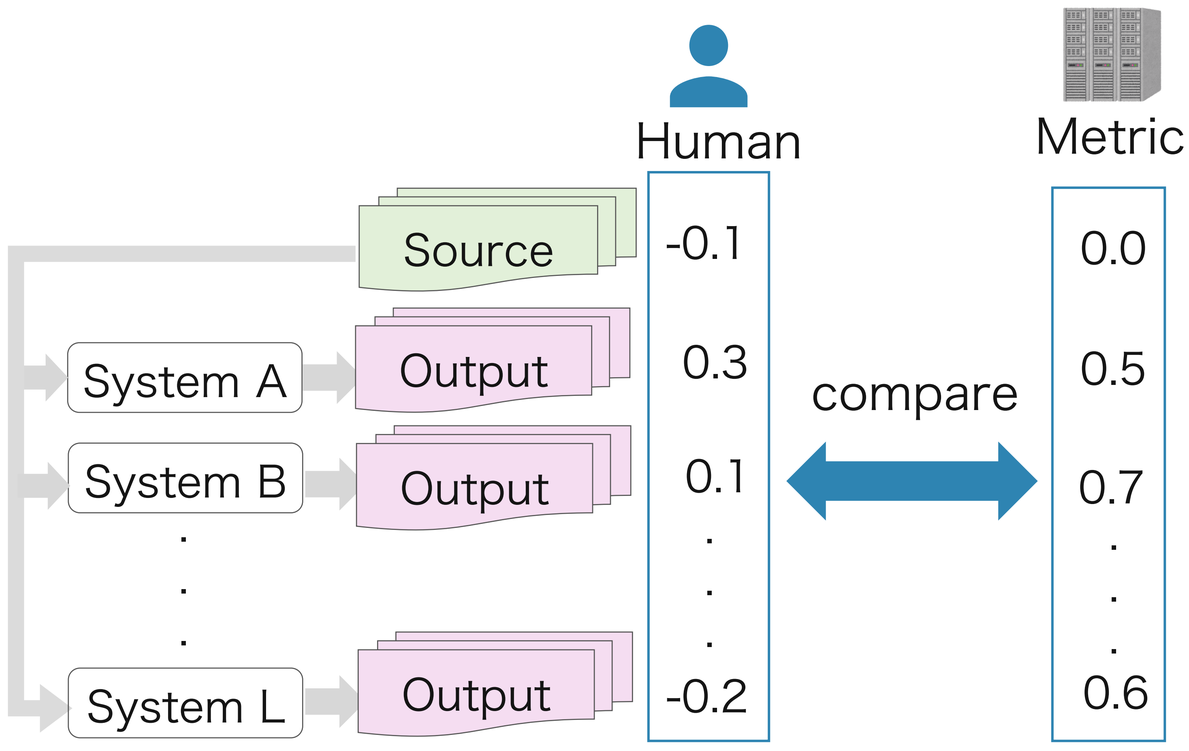
h_c, s_c: content words in h, s

The above three criteria are combined as follows:

$$Score(h,s) = \alpha S_G(h) + \beta S_F(h) + \gamma S_M(h,s)$$

$$where\ \alpha + \beta + \gamma = 1, S_G, S_F, \text{and } S_M \text{ are } [0,1]$$

Experiment: Scoring GEC systems



- We compared the proposed reference-less metric with respect to how closely each metric correlates with human ratings.
- We used the human ratings of the 12 GEC systems submitted to the CoNLL-2014 Shared Task on GEC, collected by Grundkiewicz et al., (2015).

Metric	Spearman's ρ
M^2(reference-based)[Dahlmeier&Ng, 2012]	0.648
GLEU+(reference-based)[Napoles et al., 2015]	0.857
Grammar	0.835
Fluency	0.819
Meaning	-0.192
Grammar+Fluency	0.819
Fluency+Meaning	0.868
Meaning+Grammar	0.813
Combination	**0.874**

- Combining the three criteria can boost the correlation with human rating.
- Meaning preservation metric exhibited poor correlation, but played a significant role when balanced with fluency metric.