SOME: Reference-less Sub-Metrics Optimized for Manual Evaluations of Grammatical Error Correction

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara and Mamoru Komachi Tokyo Metropolitan University, yoshimura-ryoma@ed.tmu.ac.jp

1. Abstract & Introduction

- Previous studies have shown that the reference-less metric of Grammatical Error Correction (GEC) is promising.
- Asano et al., 2017 achieved higher performance than reference-based metrics by integrating sub-metrics of three perspectives of grammaticality, fluency, and meaning preservation.
- However, each sub-metric is not optimized for manual evaluation of the system output.
  - There is no dataset of system output with manual evaluation, which is ideal for training the metric.
  - There is still room for improvement.

➢ We manually evaluated the output of GEC systems to optimize the metric.

➢ We proposed a reference-less metric trained on the created dataset.

➢ Experimental results showed that the proposed metric improves the correlation with manual evaluation in both system-and sentence-level meta-evaluation.

2. Proposed Method

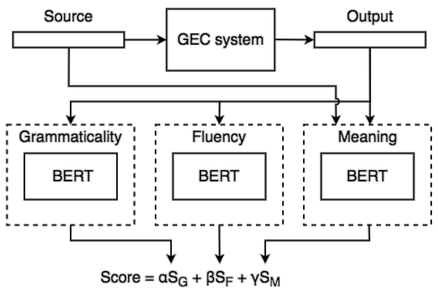
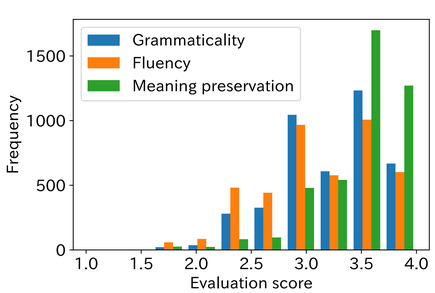


Fig1. Overview

- We integrate sub-metrics of the three perspectives.
- BERT (Devlin et al., 2019) is used for each sub-metric and fine-tuned with the created data.
- The final score is calculated using the weighted sum of each score following Asano et al., 2017.
- S_G, S_F, and S_M are normalized scores of each sub-metric.
- The non-negative weights satisfy $\alpha + \beta + \gamma = 1$.

3. Manual Evaluation of GEC System Outputs

- We collected manual evaluations of the typical five system output from CoNLL2013.
- We used Amazon Mechanical Turk and created 4,221 sentences.



Source text: This will *inversely* improve the *sale* of the shop.
System output: This will *definitely* improve the *sales* of the shop.
Grammaticaly: 3.8 **Fluency:** 3.8 **Meaning:** 1.6

Source text: The *increasing* longevity is due to fast development of *the* society so as the living pressure.
System output: The *increase* in longevity is due to *the* fast development of society so as the living pressure.
Grammaticaly: 2.6 **Fluency:** 2.4 **Meaning:** 3.8

Fig2. Histogram of each manual evaluation and examples of annotation.

4. Experimental Settings

- Our dataset was divided into train/dev/test with 3,376/422/423
- We used a publicly available pre-trained $BERT_{BASE}$ cased model.
- GUG data (Heilman et al., 2014), Lau et al., 2014, and STS dataset (Cer et al.,2015) for existing data.
- System-level meta-evaluation
  - Grundkiewicz et al., 2015
  - Correlation with a manual ranking of 12 systems.
  - The weights are tuned on the JFLEG dataset.
- Sentence-level meta-evaluation
  - Grundkiewicz et al., 2015
  - Evaluate the pairs of ranked two sentences
  - The dataset is divided. (1:9 for dev:test)

5. Experimental Results

	System-level			Sentence-level		
	Pearson	Spearman	Weights (α:β:γ)	Pearson	Spearman	Weights (α:β:γ)
M^2	0.674	0.720	-	0.464	0.294	-
GLEU	0.846	0.186	-	0.670	0.354	-
Asano et al. (2017)	0.878	0.874	0.07:0.83:0.10	0.690	0.390	0.02:0.82:0.16
SOME (BERT w/ existing data)	0.939	0.929	0.84:0.01:0.15	0.744	0.502	0.86:0.13:0.01
SOME (BERT w/ our data)	**0.975**	**0.978**	0.01:0.98:0.01	**0.749**	**0.510**	0.55:0.43:0.02

Table 1. Meta-evaluation of reference-based metrics (upper) and reference-less metrics (lower) .

		Our data		Grundkiewicz et al., 2015			
		Sentence-level		System-level		Sentence-level	
	Perspective	Pearson	Spearman	Pearson	Spearman	Accuracy	Kendall
Asano et al. (2017)	Grammaticality	0.342	0.358	0.759	0.835	0.641	0.283
	Fluency	0.220	0.238	0.864	0.819	0.707	0.415
	Meaning	0.593	0.504	**0.198**	**−0.192**	0.189	**0.059**
SOME (BERT w/ existing data)	Grammaticality	0.608	0.624	0.966	0.967	0.735	0.483
	Fluency	0.545	0.548	0.865	0.742	0.714	0.443
	Meaning	0.570	0.355	−0.462	−0.610	0.502	0.016
SOME (BERT w/ our data)	Grammaticality	**0.700**	**0.719**	**0.976**	**0.973**	**0.745**	**0.502**
	Fluency	**0.676**	**0.696**	**0.979**	**0.978**	**0.741**	**0.494**
	Meaning	**0.639**	**0.619**	−0.517	−0.621	**0.504**	0.022

Table 2. Intrinsic (Our data) and extrinsic (Grundkiewicz) meta-evaluation of each sub-metric.

6. Example

Source	There are a lot of disadvantages that people may not realize of .				
Reference	There are a lot of disadvantages that people may not realize .				
Corrected Sentence 1	There are a lot of **problems** that people may not realize .				
	Manual evaluation	M^2	GLEU	Asano et al. (2017)	SOME
	✓	0.556	0.586	0.949	**0.913**
Corrected Sentence 2	There are a lot of **the** disadvantages that people may not realize .				
	Manual evaluation	M^2	GLEU	Asano et al. (2017)	SOME
	×	0.556	0.630	0.977	0.826

Table 3. Example showing that our proposed metric works well.