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, Tokyo, Japan,
{yoshimura-ryoma, kaneko-masahiro}@ed.tmu.ac.jp
†Osaka University, Osaka, Japan, kajiwara@ids.osaka-u.ac.jp
‡Tokyo Metropolitan University, Tokyo, Japan, komachi@tmu.ac.jp

Abstract

We propose a reference-less metric trained on manual evaluations of system outputs for grammatical error correction. Previous studies have shown that reference-less metrics are promising; however, existing metrics are not optimized for manual evaluation of the system output because there is no dataset of system output with manual evaluation. This study manually evaluates the output of grammatical error correction systems to optimize the metrics. Experimental results show that the proposed metric improves the correlation with manual evaluation in both system- and sentence-level meta-evaluation. Our dataset and metric will be made publicly available.[1]

1 Introduction

Grammatical error correction (GEC) is the task of automatically correcting grammatically incorrect sentences, especially those written by language learners. To develop GEC systems efficiently, we construct an evaluation metric that has a high correlation with manual evaluations.

Reference-based metrics such as Max Match (M^2) (Dahlmeier and Ng, 2012) and GLEU (Napoles et al., 2015) are commonly used for automatic evaluation in the GEC task. However, these metrics penalize sentences whose words or phrases are not included in the reference, even if they are correct expressions because it is difficult to cover all possible references (Choshen and Abend, 2018). In contrast, reference-less metrics (Napoles et al., 2016; Asano et al., 2017) do not suffer from this limitation. Among them, Asano et al. (2017) achieved a higher correlation with manual evaluations than reference-based metrics by integrating sub-metrics from the three perspectives of (i) grammaticality, (ii) fluency, and (iii) meaning preservation. However, the correlation with the manual evaluation of system output can be further improved because they are not considered for optimizing each sub-metric.

To achieve a better correlation with manual evaluation, we create a dataset to optimize each sub-metric to the manual evaluation of GEC systems. Our annotators evaluated the output of five typical GEC systems in terms of each sub-metric of Asano et al. (2017). We propose a reference-less metric consisting of sub-metrics that are optimized for manual evaluation (SOME). It combines three regression models trained on our dataset. Experimental results show that the proposed metric improves correlation with the manual evaluation in both system- and sentence-level meta-evaluation. Detailed analysis reveals that optimization for both the manual evaluation and the output of GEC systems contribute to improvement.

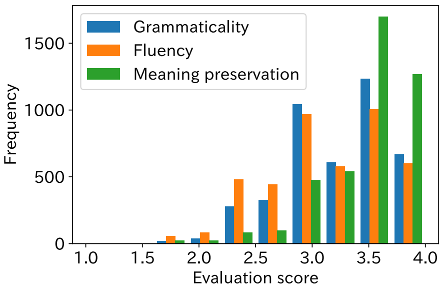
2 Related Work

Napoles et al. (2016) pioneered the reference-less GEC metric. They presented a metric based on grammatical error detection tools and linguistic features such as language models, and demonstrated that its performance was close to that of reference-based metrics. Asano et al. (2017) combined three sub-metrics: grammaticality, fluency, and meaning preservation, and outperformed reference-based metrics. They trained a logistic regression model on the GUG dataset[2] (Heilman et al., 2014) for the sub-metric

[1] https://github.com/kokeman/SOME

[2] https://github.com/EducationalTestingService/gug-data

Proceedings of the 28th International Conference on Computational Linguistics, pages 6516–6522
Barcelona, Spain (Online), December 8-13, 2020



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

Figure 1: Histogram of each manual evaluation and examples of annotation.

of grammaticality. Although the GUG is a dataset annotated for grammaticality to sentences written by language learners, our target is the learner sentence corrected by the GEC system. They used a language model and METEOR (Denkowski and Lavie, 2014) as sub-metric for fluency and meaning preservation, respectively; yet these sub-metrics are not optimized for manual evaluation. The weighted linear sum of each evaluation score was used as the final score. Although our metric follows Asano et al. (2017), each sub-metric is trained on our dataset, thus achieving a higher correlation with manual evaluation.

Apart from the GUG dataset, a fluency annotated dataset[3] (Lau et al., 2015) exists with manual evaluations of acceptability for pseudo-error sentences generated by round-trip translation of English sentences from the British National Corpus (BNC) and Wikipedia using Google Translate. We assume that sentences written by learners or translated by systems have different properties from those generated by GEC systems, and thus we collected manual evaluations of the output of the GEC systems to train the metrics. In this study, these datasets are referred to as existing data.

3 Manual Evaluation of GEC System Outputs

Data and GEC systems We collected manual evaluations for the grammaticality, fluency, and meaning preservation of the system outputs of 1,381 sentences from CoNLL 2013,[4] which are often used to evaluate GEC systems. To collect the manual evaluations for various system outputs, each source sentence was corrected by the following five typical systems: statistical machine translation (**SMT**) (Grundkiewicz and Junczys-Dowmunt, 2018), recurrent neural network (**RNN**) (Luong et al., 2015), convolutional neural network (**CNN**) (Chollampatt and Ng, 2018), self-attention network (**SAN**) (Vaswani et al., 2017), and SAN with copy mechanism (**SAN+Copy**) (Zhao et al., 2019). More details can be found in Appendix A.

Annotation By excluding duplicate corrected sentences, manual evaluation for the grammaticality, fluency, and meaning preservation were assigned to a total of 4,223 sentences, as follows: **Grammaticality**: Annotators evaluated the grammatical correctness of the system output. We followed the five-point scale evaluation criteria (4: Perfect, 3: Comprehensible, 2: Somewhat comprehensible, 1: Incomprehensible, and 0: Other) proposed by Heilman et al. (2014). **Fluency**: Annotators evaluated how natural the sentence sounds for native speakers. We followed the criteria (4: Extremely natural, 3: Somewhat natural, 2: Somewhat unnatural, and 1: Extremely unnatural) proposed by Lau et al. (2015). **Meaning preservation**: Annotators evaluated the extent to which the meaning of source sentences is preserved in system output. We followed the criteria (4: Identical, 3: Minor differences, 2: Moderate differences, 1: Substantially different, and 0: Other) proposed by Xu et al. (2016). We used Amazon Mechanical Turk[5] and recruited five native English annotators. The average of the ratings excluding "0: Other" was used as the final sentence score. For more details, refer to Appendix B. Finally, we created a dataset with manual evaluations for a total of 4,221 sentences, excluding sentences in which three or more annotators

[3] https://clasp.gu.se/about/people/shalom-lappin/smog/experiments-and-datasets

[4] https://www.comp.nus.edu.sg/~nlp/conll13st.html

[5] https://www.mturk.com/

answered “0: Other.”[6] Figure 1 presents a histogram of manual evaluations and examples of annotation. Ratings of 2 or lower generally exhibited a low frequency; the majority of the meaning preservation ratings were 3 or higher.

4 Automatic Evaluation of GEC using BERT

Using our dataset introduced in the previous section, we trained regression models corresponding to each sub-metric of (Asano et al., 2017). For grammaticality and fluency, the manual evaluations were estimated only from the system outputs, whereas, for meaning preservation, the manual evaluations were estimated from pairs of source sentences and system outputs. We used BERT (Devlin et al., 2019) for the regression models. BERT is a sentence encoder pre-trained with large-scale corpora, such as Wikipedia, based on both masked language modeling and next sentence prediction, which can achieve high performance in various natural language processing tasks by fine-tuning on a small dataset of the target task. We fine-tuned three BERT models for each perspective of grammaticality, fluency, and meaning preservation, and constructed sub-metrics that were optimized for manual evaluations of each perspective.

The final evaluation score is calculated by the weighted linear sum of each evaluation score: $\mathrm{Score} = \alpha \cdot \mathrm{Score_G} + \beta \cdot \mathrm{Score_F} + \gamma \cdot \mathrm{Score_M}$, following Asano et al. (2017). $\mathrm{Score_G}$, $\mathrm{Score_F}$, and $\mathrm{Score_M}$ are the normalized scores for grammaticality, fluency, and meaning preservation, respectively. The non-negative weights satisfy $\alpha + \beta + \gamma = 1$.

5 Experimental Setting

To verify the effectiveness of the proposed metric (SOME), we performed system- and sentence-level meta-evaluation and compared the results with those of existing metrics. Furthermore, to verify the effectiveness of our dataset based on the GEC systems, we compared our metric with a BERT-based metric fine-tuned on datasets not based on the GEC systems.

5.1 Fine-tuning BERT

SOME (BERT w/ existing data) The existing datasets described in Section 2 were used for grammaticality[2] and fluency[3] sub-metrics for fine-tuning BERT in the baseline method. For fluency, the dataset from the BNC was used for training the fluency, whereas the dataset from Wikipedia was used for development. For meaning preservation, we used the dataset[7] of the Semantic Textual Similarity task (Cer et al., 2017), which evaluates the semantic similarity between two sentences using continuous values in $[0.0, 5.0]$.

SOME (BERT w/ our data) Our dataset, introduced in Section 3, was divided into train/dev/test with 3,376/422/423 sentences and used for fine-tuning BERT[8], hyperparameter tuning, and intrinsic evaluation of each sub-metric, respectively. Refer to Appendix C for the hyperparameter settings.

5.2 Meta-Evaluation

System-level meta-evaluation In the system level meta-evaluation, the average of the sentence scores was used as the system score , and the correlation coefficients with the manual evaluations were calculated. Following Asano et al. (2017), system-level meta-evaluation was performed using Pearson’s correlation coefficient and Spearman’s rank correlation coefficient with the manual ranking of 12 systems described in (Grundkiewicz et al., 2015). The weights of the evaluation score (α, β, and γ) were tuned on the JFLEG dataset (Napoles et al., 2017), following Asano et al. (2017). To perform a comprehensive evaluation considering all perspectives, we performed a grid search in increments of 0.01 in the range of 0.01 to 0.98 for each weight and maximized Pearson’s correlation coefficient. Following the recommendation of Graham and Baldwin (2014), we used Williams significance test to identify differences in correlation that are statistically significant.

[6] Incomplete or unclear sentences.

[7] `http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark`

[8] `https://github.com/huggingface/transformers`, (BERT-BASE-CASED)

	System-level			Sentence-level		
	Pearson	Spearman	Weights (α:β:γ)	Accuracy	Kendall	Weights (α:β:γ)
M^2	0.674	0.720	-	0.464	0.294	-
GLEU	0.846	0.816	-	0.670	0.354	-
Asano et al. (2017)[9]	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 references-less metrics (lower). $*$ indicates significant difference ($p < 0.05$) between SOME (BERT w/ existing data) and SOME (BERT w/ our data). † indicates significant difference ($p < 0.05$) between Asano et al. (2017) and SOME metrics. (It was not calculated at the system-level because the scores of Asano et al. (2017) at the system-level are cited from the paper.)

		Our data		(Grundkiewicz et al., 2015)			
		Sentence-level		System-level		Sentence-level	
	Perspective	Pearson	Spearman	Pearson	Spearman	Accuracy	Kendall
Asano et al. (2017)[9]	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 (left) and extrinsic (right) meta-evaluation of each sub-metric.

Sentence-level meta-evaluation In the sentence level meta-evaluation, we used the dataset described in Grundkiewicz et al. (2015). In this dataset, output sentences from multiple GEC systems for the same input sentences are ranked by overall manual evaluation. We determined the superiority or inferiority of any two output sentences and evaluated the accuracy and Kendall's rank correlation coefficient τ. Note that sentence pairs with the same ranking were not used. The weights of the evaluation score (α, β, and γ) were tuned by dividing the dataset (Grundkiewicz et al., 2015) for development set and test set at a ratio of 1:9 and maximizing Kendall's rank correlation coefficient in the development set. The grid search range was the same as in the system-level meta-evaluation. For the significance test, we used bootstrap resampling significance tests (Graham et al., 2014).

6 Results and Discussion

Table 1 presents the results of the system- and sentence-level meta-evaluations. As the metrics based on BERT performed much better than the other metrics, the effectiveness of optimizing the sub-metrics based on the pre-trained language model for the manual evaluations of system output was confirmed. The difference in the datasets used for BERT fine-tuning indicated that using our dataset achieved a higher correlation with the manual evaluations in both the system- and sentence-level meta-evaluations. The weight of meaning preservation is small overall. We think this is because, in GEC, many common words exist between the source sentence and corrected sentence, so that, in many cases, the meaning does not change regardless of whether the correction is good or bad. The higher fluency weight for Asano et al. (2017) and SOME (BERT w/ our data) at the system level is considered to be because JFLEG, which emphasizes fluency, was used for tuning the weight. We believe that the reason why the higher gram-

[9] The system-level results are cited from the paper; the others are the results of our re-implementation.

Source sentence	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.

Source sentence	Therefore I believe the parents have their right to know the healthiness of their child .				
Reference	Therefore , I believe the parents have the right to know about the healthiness of their child .				
Corrected sentence 1	Therefore , I believe parents have their right to know the healthiness of their child .				
	Manual evaluation	M^2	GLEU	Asano et al. (2017)	SOME
	✓	0.456	0.320	0.850	0.873
Corrected sentence 2	Therefore , I believe parents have their right to know the healthiness of their *children* .				
	Manual evaluation	M^2	GLEU	Asano et al. (2017)	SOME
	×	0.333	0.245	**0.883**	**0.881**

Table 4: Example where reference-less metrics do not work properly.

maticality weight of SOME (BERT w/ existing data) at the system level is because the grammaticality sub-metric is more correlated with the JFLEG dataset than the fluency sub-metric. (Pearson's correlation coefficient of 0.963 for the grammaticality sub-metric, and 0.957 for the fluency sub-metric.)

Table 2 presents the results of each sub-metric meta-evaluation on our data (intrinsic) and the dataset described in Grundkiewicz et al. (2015) (extrinsic). In the extrinsic meta-evaluation, the grammaticality and fluency sub-metrics outperformed the baseline metrics, but the meaning preservation sub-metric did not have positive correlation. Note that in the intrinsic meta-evaluation, the correlation of each sub-metric was calculated with the manual evaluations corresponding to each perspective, whereas, in the extrinsic meta-evaluation, it was calculated with comprehensive human ranking.

7 Examples

We compared each metric for the evaluation data of Grundkiewicz et al. (2015). Table 3 shows an example where SOME (BERT w/ our data) works well. GLEU underestimates corrected sentence 1 because it does not contain "problems" in the reference sentence and overestimates corrected sentence 2, which is superficially similar but contains a superfluous "the." Conversely, SOME can make an appropriate evaluation independent of the superficial word match. Table 4 shows an example of reference-less metrics that do not work properly. In the reference-based metrics, because the reference contains "child," the corrected sentence 2 containing "child" is highly evaluated. Conversely, in the reference-less metrics, the corrected sentence 2, in which the "child" part has become "children," is highly evaluated.

8 Conclusions

We created a dataset with the manual evaluations of grammaticality, fluency, and meaning preservation for GEC system output and proposed a BERT-based reference-less metric, SOME, in which each sub-metric was optimized for each manual evaluation. The experiments demonstrated that the proposed metric achieved the highest correlation with the manual evaluations in both the system- and sentence-level meta-evaluations. Furthermore, the effectiveness of optimizing the metrics for manual evaluations on the GEC system output was confirmed by comparison with BERT fine-tuned on existing datasets.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP20K19861. We would like to thank Hiroki Asano for giving the implementation code and Keisuke Sakaguchi for the system output of JFLEG.

References

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Leshem Choshen and Omri Abend. 2018. Inherent biases in reference-based evaluation for grammatical error correction and text simplification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176. Association for Computational Linguistics.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274. Association for Computational Linguistics.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2018. Near human-level performance in grammatical error correction with hybrid machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 284–290.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1618–1628.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Masato Mita, Tomoya Mizumoto, Masahiro Kaneko, Ryo Nagata, and Kentaro Inui. 2019. Cross-corpora evaluation and analysis of grammatical error correction models — is single-corpus evaluation enough? In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1309–1314.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234.

Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165.