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

Hiroki Asano[1][2], **Tomoya Mizumoto**[2], and **Kentaro Inui**[1][2]
[1] Graduate School of Information Sciences, Tohoku University
[2] RIKEN Center for Advanced Intelligence Project
asano@ecei.tohoku.ac.jp,tomoya.mizumoto@riken.jp
inui@ecei.tohoku.ac.jp

Abstract

In grammatical error correction (GEC), automatically evaluating system outputs requires gold-standard references, which must be created manually and thus tend to be both expensive and limited in coverage. To address this problem, a reference-less approach has recently emerged; however, previous reference-less metrics that only consider the criterion of grammaticality, have not worked as well as reference-based metrics. This study explores the potential of extending a prior grammaticality-based method to establish a reference-less evaluation method for GEC systems. Further, we empirically show that a reference-less metric that combines fluency and meaning preservation with grammaticality provides a better estimate of manual scores than that of commonly used reference-based metrics. To our knowledge, this is the first study that provides empirical evidence that a reference-less metric can replace reference-based metrics in evaluating GEC systems.

1 Introduction

Grammatical error correction (GEC) has been an active research area since a series of shared tasks was launched at CoNLL (Ng et al., 2013, 2014). The GEC mainly constitutes a generative task, i.e., a task that produces a grammatically correct sentence from a given original sentence whereby multiple distinct outputs can be judged "correct" for a single input. Therefore, automatically evaluating the performance is not straightforward and is considered as an important issue as in the fields of translation and summarization.

A common approach to automatically evaluating GEC systems involves reference-based evaluation, where gold-standard references are manually created for a given test set of original sentences and each system output is scored by comparing it with corresponding gold-standard references with some metrics (referenced-based metric) (Dahlmeier and Ng, 2012; Felice and Briscoe, 2015; Napoles et al., 2015), analogous to BLEU (Papineni et al., 2002) in machine translation. Reference-based evaluation, however, has a severe drawback. In GEC, multiple outputs can be a right answer for a single input sentence. If the gold-standard references at hand lack coverage, reference-based metrics may unfairly underestimate system performance. One way to cope with this problem is to exhaustively collect potential corrections; however, this is not straightforward and can be of immense cost.

As an alternative approach to this problem, Napoles et al. (2016) proposed a new method that does not require gold-standard references (i.e., *reference-less* metric). Their idea is to evaluate GEC system performance by assessing the grammaticality of system outputs without gold-standard references. This approach is advantageous in that it does not require manual creation of references. The results of the experiments reported in (Napoles et al., 2016), however, reveal that their reference-less metric cannot evaluate GEC systems as well as a reference-based metric, GLEU+ (Napoles et al., 2015).

Given the above, we explore the potential capabilities of reference-less evaluation by extending grammaticality-based method of Napoles et al. (2016) with other assessment criteria. More specifically, we consider the criteria of fluency and meaning preservation as additions to grammaticality and empirically show that a reference-less metric that combining these three criteria can evaluate

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 343–348,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

GEC systems better than reference-based metrics. To our best knowledge, this is the first study that provides such empirical evidence to show that a reference-less metric can replace reference-based metrics in evaluating GEC systems.

2 Reference-less GEC assessment

There are two key ideas behind our reference-less approach to GEC assessment. First, we explore a range of criteria for assessing grammatical corrections that are considered important in the GEC literature and can be automated without reference data. Second, we identify a system that combines the aforementioned criteria to provide a better estimate of manual scores as compared with reference-based metrics. Given these two key ideas, we consider the following three criteria: grammaticality, fluency, and meaning preservation.

Grammaticality The criterion of grammaticality in the metric defined by Napoles et al. (2016) is modeled on the linguistic-feature-based model originally proposed by Heilman et al. (2014). We also use a similar method. More specifically, for a hypothesis h, the grammaticality score $\mathrm{S_G}(h)$ is determined by a logistic regression with linguistic features, including the number of misspellings, language model scores, out-of-vocabulary counts, and PCFG and link grammar features. We extend this model by incorporating the number of errors detected by the Language Tool[1]. Further, we trained our model using the GUG dataset (Heilman et al., 2014) and the implementation provided by Napoles et al. (2016).[2] In addition, we used Gigaword (Parker et al., 2011) and TOEFL11 (Blanchard et al., 2013) to train the language model. The resulting grammaticality model achieved an accuracy of 78.9%, slightly higher than the original model (77.2%), in the binary prediction of grammaticality on the GUG dataset.

Fluency The importance of fluency in GEC has been shown by Sakaguchi et al. (2016) and Napoles et al. (2017), however there are no evaluation metrics that consider fluency. Fluency can be captured by statistical language modeling (Lau et al., 2015). More specifically, for a hypothesis h, fluency score $\mathrm{S_F}(h)$ is calculated as follows:[3]

$$\mathrm{S_F}(h) = \frac{\log P_m(h) - \log P_u(h)}{|h|}, \qquad (1)$$

where $|h|$ denotes the sentence length, $P_m(h)$ denotes the probability of the sentence given by a language model, and $P_u(h)$ denotes the unigram probability of the sentence. In our study, we adopted Recurrent Neural Network Language Models implemented via faster-rnnlm.[4] Further, we used 10 million sentences from the British National Corpus (BNC Consortium, 2007) and Wikipedia. Given these datasets, we found the fluency scored by our model to have a correlation coefficient (Pearson's r) of 0.395 with acceptability scored by humans in the same setting described by Lau et al. (2015).

Meaning preservation In GEC, the meaning of original sentences should be preserved. As an example, consider sentence (1a) below being revised to form sentence (1b).

(1) a. *It is unfair to release a law only point to the genetic disorder.* (original)

b. *It is unfair to pass a law.* (revised)

Sentence (1b) is grammatically correct, but does not preserve the meaning of sentence (1a), and thus sentence (1b) should be considered as inappropriate. To assess how much of the meaning of an original sentence is preserved in a revision, one can consider the use of an evaluation metric devised in the MT field. In this study, we adopt METEOR (Denkowski and Lavie, 2014) because it focuses on semantic similarity much more so than other common metrics, such as BLEU. Meaning score $\mathrm{S_M}(h, s)$ for input of a source sentence s and a hypothesis h is calculated as follows:

$$\mathrm{S_M}(h, s) = \frac{P \cdot R}{t \cdot P + (1-t) \cdot R}, \qquad (2)$$

where $P = \frac{m(h_c, s_c)}{|h_c|}$ and $R = \frac{m(h_c, s_c)}{|s_c|}$. h_c denotes content words in the hypothesis h, s_c denotes content words in the source sentence s, and $m(h_c, s_c)$ denotes the number of matched content words between the output and the original sentence, and

[1] https://languagetool.org

[2] https://github.com/cnap/grammaticality-metrics/tree/master/heilman-et-al

[3] In many cases $\mathrm{S_F}(h)$ is more than 0 and less than 1. When it is less than 0, $\mathrm{S_F}(h) = 0$, and when it is more than 1, $\mathrm{S_F}(h) = 1$

[4] https://github.com/yandex/faster-rnnlm

is calculated considering inflection, synonyms and misspellings[5]. Note that we use $t = 0.85$, which is a default value provided of METEOR.

The above three criteria are combined as follows:

$$\text{Score}(h, s) = \alpha \text{S}_\text{G}(h) + \beta \text{S}_\text{F}(h) + \gamma \text{S}_\text{M}(h, s), \quad (3)$$

where the ranges of S_G, S_F, and S_M are [0, 1] and $\alpha + \beta + \gamma = 1$. We choose these weights empirically with a development dataset.

3 Experiments

We conducted two experiments to investigate the extent to which our reference-less metric is close to human evaluation compared with baseline reference-based metrics. We used two commonly used reference-based metrics M^2 (Dahlmeier and Ng, 2012) and GLEU+ (Sakaguchi et al., 2016; Napoles et al., 2016) (A modified version of GLEU (Napoles et al., 2015)).

3.1 Automatic ranking of GEC systems

We first compare the proposed reference-less metric with respect to how closely each metric correlates with human ratings.

For this experiment, we used the CoNLL-2014 Shared Task (CoNLL) dataset (Ng et al., 2014). The CoNLL dataset is a collection of the outputs produced by the 12 participant GEC systems submitted to the CoNLL-2014 Shared Task, where the 12 GEC systems' outputs to each input student sentence are ranked by multiple human raters (Grundkiewicz et al., 2015). An advantage of using this dataset is that it includes an extensive set of references for each input student sentence: two references originally provided in the CoNLL-2014 Shared Task, eight references provided by Bryant and Ng (2015), and eight references provided by Sakaguchi et al. (2016). In the experiment, we used all the 18 references for the baseline reference-based metrics in order to bring out the maximal potential of those metrics.

For tuning the weights, α, β and γ, of our metric, we used another distinct dataset, the JHU FLuency-Extended GUG (henceforth, JFLEG) dataset (Napoles et al., 2017). This dataset is a collection of tuples of an input student sentence, four GEC system outputs, and a human rating. We selected weights with the highest correlation on the JFLEG dataset, obtaining $\alpha = 0.07$, $\beta = 0.83$, and $\gamma = 0.10$. Note that these optimized weights should not be interpreted as the relative importance of the subcomponents because outputs of those subcomponents differ in variance.

For testing, following the experiments reported in (Napoles et al., 2016), the 12 system outputs for each input student sentence were scored with each metric, and next for each metric, the 12 systems were ranked according to their averaged scores. Each metric's ranking was then compared to the human ranking of Grundkiewicz et al. (2015, Table 3c[6]) to compute the correlation coefficients, Spearman's ρ and Pearson's r.

The results are shown in Table 1. Many interesting findings can be drawn. The grammaticality metric alone, which corresponds to (Napoles et al., 2016), outperformed M^2 but did not perform as well as GLEU+. The meaning preservation metric exhibited poor correlation with human ranking; however, when combining meaning preservation with fluency, the prediction capability boosted, prevailing over GLEU+. We believe this result makes good sense because the meaning preservation metric, i.e. METEOR, relies mostly on shallow similarity (although it partially considers paraphrases) and tends to prefer system outputs with fewer corrections; nevertheless, it plays a significant role when balanced with fluency. Combining all the three subcomponents even further improved Spearman's ρ ($\rho = 0.874$), significantly outperforming both M^2 and GLEU+. To our knowledge, this is the first study that provides empirical evidence that a reference-less metric can correlate better with human ratings compared with the state-of-the-art reference-based metrics in evaluating GEC systems.

3.2 Minimal edits vs. fluent edits

According to recent work by Sakaguchi et al. (2016), the aspect of fluency is potentially even further important than ever considered in the GEC literature. We expect that this emphasis on fluency might bring further advantages to reference-less metrics as opposed to reference-based metrics.

[5]In order to handle misspellings, we first ran a spell checker on a given input sentence to obtain candidate corrections and then put them into METEOR to find the maximum score.

[6]We used the TrueSkill ranking simply because (i) we wanted to compare our results with those reported in Napoles et al. (2016), where only TrueSkill was used, and (ii) system outputs in the JFLEG are also ranked with TrueSkill.

Metric	Spearman's ρ	Pearson's r
M^2	0.648	0.632
GLEU+	0.857	0.843
Grammar	0.835	0.759
Meaning	-0.192	0.198
Fluency	0.819	0.864
Grammar+Meaning	0.813	0.794
Meaning+Fluency	0.868	0.876
Fluency+Grammar	0.819	0.864
Combination (proposed)	**0.874**	**0.878**

Table 1: Correlation between human and metric rankings.

In their recent work, Sakaguchi et al. (2016) created an interesting dataset by asking four human editors to produce one *minimal edit* (minimal grammatical error corrections) and one *fluent edit* (extended corrections with maximal fluency) for each original student sentence in the aforementioned CoNLL dataset (corresponding to the "eight references provided by Sakaguchi et al. (2016)" referred to in 3.1). Using this dataset, Sakaguchi et al. showed that human raters clearly prefer fluency edits to minimal edits.

An intriguing question here is whether our reference-less metric (the combination of grammaticality, fluency and meaning-preservation) is indeed capable of preferring fluent edits to minimal edits despite that fluent edits are less similar to their original sentences than minimal edits. We therefore conducted a supplemental experiment as follows.

We chose two editors out of the four editors employed for Sakaguchi et al. (2016)'s dataset and extracted the four edits by these two editors (Editor A and Editor B) for each original student sentence, fluent edits by Editor A (Flu-A), minimal edits by Editor A (Min-A), fluent edits by Editor B (Flu-B), minimal edits by Editor B (Min-B), as the test set. We then applied our metric and the two baseline metrics to this test set to rank the four sets of edits, where the reference-based metrics used the remaining references (the 14 references for each original sentence).

The results are shown in Table 2. While the proposed metric (Combination) consistently prefers fluent edits, the reference-based metrics seriously underestimate fluent edits. GLEU+ consistently preferred the minimal edits. This is somewhat an expected result because the majority of the reference data consists of "minimal edits" reflecting the nature of the GEC task and GLEU+ tends to lean towards the majority of the references. One straightforward way to cope with this problem is to collect as many diverse fluent edits as possible, which would be prohibitively costly though. M^2 may not suffer the same problem; however, as revealed by our first experiment, M^2 can be far less appropriate as a metric for GEC assessment compared with GLEU+ (see Table 1). In contrast, the proposed reference-less approach has good potential for this issue. Table 3 shows an example where the proposed metric preferred a fluent edit but the reference-based metrics preferred a minimal edit. The reference-based metrics gave low scores to the fluent edit because the human references did not cover the correction "made the world smaller".

rank	Combination (proposed)	M^2	GLEU+
1	Flu-A (0.865)	Min-B (0.641)	Min-B (0.628)
2	Flu-B (0.854)	Flu-A (0.634)	Flu-B (0.607)
3	Min-B (0.848)	Flu-B (0.626)	Min-A (0.606)
4	Min-A (0.844)	Min-A (0.590)	Flu-A (0.563)

Table 2: Rankings of the four reference sets with each metric. Scores assigned by the GEC metrics are shown in parentheses.

Sentence	Comb.	M^2	GLEU+
From this scope, social media has shortened our distance. (minimal)	0.541	1.00	0.575
From this perspective, social media has made the world smaller. (fluent)	0.688	0.277	0.251

Table 3: An example that the reference-less metric works well and the sentence-level scores by GEC metrics.

4 Discussion and Conclusions

In this paper, we have presented a reference-less approach to automatic assessment of GEC systems and have empirically shown that combining the three criteria of grammaticality, fluency, and meaning preservation can boost the correlation with human ratings. To our best knowledge, the paper has provided the first empirical evidence supporting the hypothesis that a reference-less metric can outperform and thus potentially replace the state-of-the-art reference-based metrics in this research field.

Our error analysis has revealed that the proposed metric still has room for improvement. One obvious point of improvement is around meaning preservation. The present choice for this component, METEOR, does not allow us to take nearly

synonymous phrases into account. For example, METEOR wrongly votes for an original sentence *our family and relatives grew us up* against a correctly revised sentence *our family and relatives brought us up*. Recent advances in computational modeling of sentence similarity (He and Lin, 2016; Rychalska et al., 2016, etc.) should be worthwhile to incorporated.

It has also turned out that the present fluency metric is undesirably affected by misspelled words. As in Equation 1, the unigram probability regularizes the sentence probability so that the score of fluency will not be underestimated by rare words. However, with misspelled words, the normalization works excessively as they are treated as rare words. This newly provides an interesting issue of how to estimate the fluency of student sentences.

Another direction for improvement is to explore methods for combining grammaticality, fluency and meaning preservation. For example, the oracle combination[7] of the three components exhibited a significantly high correlation with the human ranking ($\rho = 0.951$, $r = 0.923$). This also indicates further room for improvement.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. *TOEFL11: A Corpus of Non-Native English*. Technical report, Educational Testing Service.

BNC Consortium. 2007. *The British National Corpus*. version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Christopher Bryant and Hwee Tou Ng. 2015. How Far are We from Fully Automatic High Quality Grammatical Error Correction? 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 Paper*, pages 697–707.

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.

Mariano Felice and Ted Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 578–587.

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.

Hua He and Jimmy J Lin. 2016. Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 937–948.

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.

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 B enchmark 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.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

[7] The weights were set to a value that achieves the highest correlation on the CoNLL dataset.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. *English Gigaword Fifth Edition LDC2011T07*. Philadelphia: Linguistic Data Consortium.

Barbara Rychalska, Katarzyna Pakulska, Krystyna Chodorowska, Wojciech Walczak, and Piotr Andruszkiewicz. 2016. Samsung Poland NLP Team at SemEval-2016 Task 1: Necessity for diversity; combining recursive autoencoders, WordNet and ensemble methods to measure semantic similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 602–608.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the Goals of Grammatical Error Correction: Fluency Instead of Grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.