

# GLEU Without Tuning

Courtney Napoles<sup>1</sup>, Keisuke Sakaguchi<sup>1</sup>, Matt Post<sup>2</sup>, and Joel Tetreault<sup>3</sup>

<sup>1</sup>Center for Language and Speech Processing, Johns Hopkins University

<sup>2</sup>Human Language Technology Center of Excellence, Johns Hopkins University

<sup>3</sup>Yahoo

## Abstract

The GLEU metric was proposed for evaluating grammatical error corrections using n-gram overlap with a set of reference sentences, as opposed to precision/recall of specific annotated errors (Napoles et al., 2015). This paper describes improvements made to the GLEU metric that address problems that arise when using an increasing number of reference sets. Unlike the originally presented metric, the modified metric does not require tuning. We recommend that this version be used instead of the original version.<sup>1</sup>

## 1 Introduction

GLEU (Generalized Language Understanding Evaluation)<sup>2</sup> was designed and developed using two sets of annotations as references, with a tunable weight to penalize n-grams that should have been changed in the system output but were left unchanged (Napoles et al., 2015). After publication, it was observed that the weight needed to be re-tuned as the number of references changed. With more references, more variations of the sentence are seen which results in a larger set of reference n-grams. Larger sets of reference n-grams tend to have higher overlap with the source n-grams, which decreases the number of n-grams that were seen in the source but not the reference. Because of this, the penalty term decreases and a large weight is needed for the penalty term to have the same magnitude as the penalty when there are fewer references.

As re-tuning the weight for different sized reference sets is undesirable, we simplified GLEU so that there is no tuning needed and the metric is

portable across comparisons against any number of references.

## 2 Modifications to GLEU

Our GLEU implementation differs from that of Napoles et al. (2015). As originally presented, in computing n-gram precision, GLEU double-counts n-grams in the reference that do not appear in the source, and it subtracts a weighted count of n-grams that appear in the source ( $S$ ) and not the reference ( $R$ ). We use a modified version, GLEU<sup>+</sup>, that simplifies this. Precision is simply the number of reference n-gram matches, minus the counts of n-grams found more often in the source than the reference (Equation 1). GLEU<sup>+</sup> follows the same intuition as the original GLEU: overlap between  $S$  and  $R$  should be rewarded and n-grams that should have been changed in  $S$  but were not should be penalized.

The precision term in Equation 1 is then used in the standard BLEU equation (Papineni et al., 2002) to get the GLEU<sup>+</sup> score. Because the number of possible reference n-grams increases as more reference sets are used, we calculate an intermediate GLEU<sup>+</sup> by randomly sample from one of the references for each sentence, and report the mean score over 500 iterations. It takes less than 30 seconds to evaluate 1,000 sentences using 500 iterations.

## 3 Results

Using this revised version of GLEU, we calculated the scores for each system submitted to the CoNLL 2014–Shared Task on Grammatical Error Correction<sup>3</sup> to update the results reported in Tables 4 and 5 of Napoles et al. (2015). The system ranking by GLEU<sup>+</sup> is compared to the originally reported GLEU (GLEU<sub>0</sub>), M<sup>2</sup>, and the human ranking (Table 1).

<sup>1</sup>Download available at <https://github.com/cnap/gec-ranking/>.

<sup>2</sup>Not to be confused with the method of the same name presented in Mutton et al. (2007).

<sup>3</sup><http://www.comp.nus.edu.sg/~nlp/conll14st.html>

$$p_n^* = \frac{\left( \sum_{ngram \in \{C \cap R\}} count_{C,R}(ngram) - \sum_{ngram \in \{C \cap S\}} \max[0, count_{C,S}(ngram) - count_{C,R}(ngram)] \right)}{\sum_{ngram \in \{C\}} count(ngram)}$$

$$count_{A,B}(ngram) = \min(\# \text{ occurrences of } ngram \text{ in A}, \# \text{ occurrences of } ngram \text{ in B})$$

Equation 1: Modified precision calculation of GLEU<sup>+</sup>.

Human	M <sup>2</sup>	GLEU <sub>0</sub>	GLEU <sup>+</sup>
CAMB	CUUI	CUUI	CAMB
AMU	CAMB	AMU	CUUI
RAC	AMU	UFC	AMU
CUUI	POST	CAMB	UMC
source	UMC	source	PKU
POST	NTHU	IITB	POST
UFC	PKU	SJTU	SJTU
SJTU	RAC	PKU	NTHU
IITB	SJTU	UMC	UFC
PKU	UFC	NTHU	IITB
UMC	IPN	POST	source
NTHU	IITB	RAC	RAC
IPN	source	IPN	IPN

Table 1: Ranking of CoNLL 2014 Shared Task system outputs, as judged by humans, M<sup>2</sup>, and both versions of GLEU.

On average, M<sup>2</sup> ranks systems within 3.4 places of the human ranking. Both GLEU scores have closer rankings on average: GLEU<sub>0</sub> within 2.6 and GLEU<sup>+</sup> within 2.9 places of the human ranking.

The correlation between the system scores and the human ranking is shown in Table 2. GLEU<sup>+</sup> has slightly stronger correlation with the human ranking than GLEU<sub>0</sub>, which is significantly greater than the human correlation with M<sup>2</sup>, however the rank correlation of GLEU<sup>+</sup> is weaker than GLEU<sub>0</sub> and M<sup>2</sup>.

## 4 Conclusion

We recommend that the originally presented GLEU no longer be used due to the issues we identified in Section 1. The updated version of GLEU that does not require tuning (GLEU<sup>+</sup>) should be used instead. The code is available at <https://github.com/cnap/gec-ranking>.

Metric	<i>r</i>	<i>ρ</i>
GLEU <sup>+</sup>	0.549	0.401
GLEU <sub>0</sub>	0.542	0.555
M <sup>2</sup>	0.358	0.429
I-measure	-0.051	-0.005
BLEU	-0.125	-0.225

Table 2: Correlation between automatic metrics and the human ranking.

## References

- Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 344–351, Prague, Czech Republic, June. Association for Computational Linguistics.
- 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, Beijing, China, July. Association for Computational Linguistics.
- 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, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.