# GLEU Without Tuning

**Courtney Napoles[1], Keisuke Sakaguchi[1], Matt Post[2],** and **Joel Tetreault[3]**
[1]Center for Language and Speech Processing, Johns Hopkins University
[2]Human Language Technology Center of Excellence, Johns Hopkins University
[3]Yahoo

## 1 Introduction

GLEU 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, we used GLEU for an increasing number of references and noticed that the weight needed to be retuned 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, for the penalty term to have the same magnitude as the penalty when there are fewer references, a large weight is needed.

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 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$^+$, that simplifies this: Precision is simply the number of reference ($R$) n-gram matches, minus the counts of n-grams found more often in the source ($S$) than the reference (Equation 1). GLEU$^+$ follows the same intuition as the original GLEU, which is that 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$^+$ score. Because the number of possible reference n-grams increases as more reference sets are used, we calculate an intermediate GLEU$^+$ 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 the revised version of GLEU, we calculated the scores for each system submitted to the CoNLL 2014–Shared Task on Grammatical Error Correction[1] to update the results reported in Napoles et al. (2015) (Tables 4 and 5). The system ranking by GLEU$^+$ is compared to the originally reported GLEU (GLEU$_0$), M$^2$, and the human ranking:

| **Human** | **M$^2$** | **GLEU$_0$** | **GLEU$^+$** |
|---|---|---|---|
| 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 |

On average, M$^2$ systems are ranked within 3.4 places of the human ranking. Both GLEU scores have average closer rankings: GLEU$_0$ within 2.6 and GLEU$^+$ within 2.9 places of the human ranking.

[1] http://www.comp.nus.edu.sg/~nlp/conll14st.html

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

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

Equation 1: Modified precision calculation of GLEU$^+$.

The correlation between the system scores and the human ranking is as follows.

| **Metric** | $r$ | $\rho$ |
|---|---|---|
| GLEU$^+$ | 0.549 | 0.401 |
| GLEU$_0$ | 0.542 | 0.555 |
| M$^2$ | 0.358 | 0.429 |
| I-measure | -0.051 | -0.005 |
| BLEU | -0.125 | -0.225 |

GLEU$^+$ has slightly stronger correlation with the human ranking than GLEU$_0$, which is significantly greater than the human correlation with M$^2$, however the rank correlation of GLEU$^+$ is weaker than GLEU$_0$ and M$^2$.

## 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$^+$) should be used instead. The code is available at `https://github.com/cnap/gec-ranking`.

## References

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.