Towards Neural Similarity Evaluators

Hassan Kané *
WL Research
hassanmohamed@alum.mit.edu

Yusuf Kocyigit *
WL Research
Boğaziçi University
yusuf.kocyigit@boun.edu.tr

Ali Abdalla
WL Research
aabdalla@alum.mit.edu

Pelkins Ajanoh
WL Research
pelkins@alum.mit.edu

Mohamed Coulibali
WL Research
Laval University
mohamed-konoufo.coulibaly.1@ulaval.ca

Abstract

We review the limitations of BLEU and ROUGE – the most popular metrics used to assess reference summaries against hypothesis summaries, and come up with criteria for what a good metric should behave like and propose concrete ways to use and test recent Transformers-based Language Models to assess reference summaries against hypothesis summaries.

1 Introduction

Evaluation metrics play a central role in the machine learning community. They direct the efforts of the research community and are used to define the state of the art models. In machine translation and summarization, the two most common metrics used for evaluating similarity between candidate and reference texts are BLEU [Papineni et al., 2002] and ROUGE [Lin, 2004]. Both approaches rely on counting the matching n-grams in the candidates summary to n-grams in the reference text. BLEU is precision focused while ROUGE is recall focused.

These metrics have posed serious limitations and have already been criticized by the academic community [Reiter, 2018] [Callison-Burch et al., 2006] [Sulem et al., 2018] [Novikova et al., 2017]. In this work, we formulate an empirical criticism of BLEU and ROUGE, establish a criteria that a sound evaluation metric should have and propose concrete ways to test any metric towards these criteria. We also use recent advances in NLP to design a data-driven metric addressing the weaknesses found in BLEU and ROUGE and scoring high on the criteria for a sound evaluation metric.

2 Related Work

2.1 BLEU, ROUGE and n-gram matching approaches

BLEU (Bilingual Evaluation Understudy) [Papineni et al., 2002] and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin, 2004] have been used to evaluate many NLP tasks for almost two decades. The general acceptance of these methods depend on many factors including

*H. Kané and Y. Kocyigit contributed equally to this work.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

their simplicity and the intuitive interpretability. Yet the main factor is the claim that they highly correlate with human judgement [Papineni et al., 2002]. This has been criticised extensively by the literature and the shortcomings of these methods have been widely studied. Reiter [Reiter, 2018] , in his structured review of BLEU, finds a low correlation between BLEU and human judgment. Callison et al [Callison-Burch et al., 2006] examines BLEU in the context of machine translation and find that BLEU does neither correlate with human judgment on adequacy(whether the hypothesis sentence adequately captures the meaning of the reference sentence) nor fluency(the quality of language in a sentence). Sulem et al [Sulem et al., 2018] examines BLEU in the context of text simplification on grammaticality, meaning preservation and simplicity and report BLEU has very low or in some cases negative correlation with human judgment.

2.2 Transformers, BERT and GPT

Language modeling has become an important NLP technique thanks to the ability to apply it to various NLP tasks as explained in Radford et al [Radford et al., 2019]. There are two leading architectures for language modeling Recurrent Neural Networks (RNNs)[Mikolov et al., 2010] and Transformers [Vaswani et al., 2017] . RNNs handle the input tokens, words or characters, one by one through time to learn the relationship between them, whereas, transformers receive a segment of tokens and learn the dependencies between them using an attention mechanism.

2.3 Model-based metrics

While BLEU and ROUGE are defined in a discrete space new evaluation metric can be defined in this continuous space. BERTscore [Zhang et al., 2019] uses word embeddings and cosine similarity to create a score array and use greedy matching to maximize the similarity score. Sentence Mover's Similarity [Clark et al., 2019] uses the mover similarity, Wasserstein distance, between sentence embedding generated from averaging the word embeddings in a sentence.

One other evaluation method proposed is RUSE [Shimanaka et al., 2018] this method proposes embedding both sentences separately and pooling them to a given size. After that they use a pre trained MLP to predict on different tasks. This quality estimator metric is then proposed to be used in language evaluation.

Our proposed methodology is to take neural language evaluation beyond architecture specifications. We are proposing a framework in which an evaluator's success can be determined.

3 Challenges with BLEU and ROUGE

In this part, we discuss three significant limitations of BLEU and ROUGE. These metrics can assign: High scores to semantically opposite translations/summaries, Low scores to semantically related translations/summaries and High scores to unintelligible translations/summaries.

3.1 High score, opposite meanings

Suppose that we have a reference summary s1. By adding a few negation terms to s1, one can create a summary s2 which is semantically opposite to s1 but yet has a high BLEU/ROUGE score.

3.2 Low score, similar meanings

In addition not to be sensitive to negation, BLEU and ROUGE score can give low scores to sentences with equivalent meaning. If s2 is a paraphrase of s1, the meaning will be the same ;however, the overlap between words in s1 and s2 will not necessarily be significant.

3.3 High score, unintelligible sentences

A third weakness of BLEU and ROUGE is that in their simplest implementations, they are insensitive to word permutation and can give very high scores to unintelligible sentences. Although higher order BLEU scores are expected to mitigate this effect, they make the metric more sensitive to paraphrasing.

	ROUGE	BLEU	RoBERTa-STS
Pearson Corr.	0.47	0.31	**0.92**

Table 1: Results of semantic similarty experiments

	ROUGE	BLEU	RoBERTa-STS
Spearman's RC	0.255	0.216	**0.744**
Kendall's τ	0.215	0.186	**0.69**

Table 2: Results of logical entailment experiments

4 Assessing evaluation metrics

4.1 Metric Scorecard

To overcome the previously highlighted challenges and provide a framework by which metrics comparing reference summaries/translation can be assessed and improved, we established first-principles criteria on what a good evaluator should do.

The first one is that it should be highly correlated with human judgement in semantic similarity. The second one is that it should be able to distinguish sentences which are in logical contradiction, logically unrelated or in logical agreement. The third one is that given s1, s2 which are semantically similar, eval(s1,s2) > eval(s1,s2(corrupted) > eval(s1,s2(more corrupted)) where corruption here includes removing words, adding noise to the word order or including grammatical mistakes.

4.2 Implementing the metric scorecard

We will now give a more detailed example to how the scorecard can be implemented. For every dimension of the scorecard the experiments are done with three metrics. BLEU with equal weights between 1 to 4 grams. ROUGE with averaging ROUGE-1 and ROUGE-2 and the a neural evaluator. The evaluator is the RoBERTa large pre-trained model [Liu et al., 2019] , which we fine tune it to predict sentence similarity (0-5 scale) on the STS-B benchmark dataset (8628 sentence pairs).

4.2.1 Semantic Similarity

The first expectation from a google similarty metric is to correlate highly with human judgment in terms of assessing semantic similarity. Here we assessed BLEU and ROUGE on the STS-B benchmark and compared their performance to a RoBERTa model fine tuned for semantic similarity (Table 1).

4.2.2 Logical Entailment

Another characteristic of a good metric is to differentiate the argument, core meaning, in a sentence and take it into account when assessing hypothesis text with references. Here we used the MNLI dataset where for each text we have three hypothesis text representing contradiction, neutral and entailment. We expect a good metric to rank entailment higher than neutral and both of them higher than contradiction. To assess the quality of a metric we propose to use the Spearman's ranked correlation and in Table 4.2.2 we also experiment with Kendall's τ. Here we observe that the RoBERTa model remarkably outperforms BLEU and ROUGE and both of these metrics show very little correlation with human judgment.

4.2.3 Robustness to Grammatical Errors

For assessing the third criteria. We start with 3479 sentence pairs from the MNLI dataset that are labelled as entailment. We introduce random corruptions such as random insertion, deletion and grammatical errors as in [Zhao et al., 2019]. We use two different set of parameters for different corruption levels and expect that a good metric would rank the original similar sentence higher than the less corrupted and both higher than the more corrupted sentence. Here we also propose to use the Spearman's ranked correlation and also experiment with Kendall's τ. We report results on Table 3. Where we see that the RoBERTa model once more outperforms BLEU and ROUGE.

	ROUGE	BLEU	RoBERTa-STS
Spearman's RC	0.528	0.472	**0.718**
Kendall's τ	0.478	0.419	**0.667**

Table 3: Results of grammatical error experiments

5 Conclusion

In this work, we have established a framework to assess metrics comparing the quality of reference and hypothesis summary/translations. Based on these criteria, we compare evaluators using recent Transformers to BLEU and ROUGE and highlight their potential to replace BLEU and ROUGE.

References

C. Callison-Burch, M. Osborne, and P. Koehn. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

E. Clark, A. Celikyilmaz, and N. A. Smith. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, 2019.

C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

J. Novikova, O. Dušek, A. C. Curry, and V. Rieser. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*, 2017.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

E. Reiter. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401, 2018.

H. Shimanaka, T. Kajiwara, and M. Komachi. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, 2018.

E. Sulem, O. Abend, and A. Rappoport. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*, 2018.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

W. Zhao, L. Wang, K. Shen, R. Jia, and J. Liu. 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, 2019.