JAUNE: JUSTIFIED AND UNIFIED NEURAL LANGUAGE EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We review the limitations of BLEU and ROUGE – the most popular metrics used to assess reference summaries against hypothesis summaries, and introduce JAUNE: a set of criteria for what a good metric should behave like and propose concrete ways to use 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 JAUNE: a set of criteria that a sound evaluation metric should pass to justify its working and propose concrete ways to use recent advances in NLP to design 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 their simplicity and intuitive interpretability. Moreover, the main factor is the claim that they highly correlate with human judgement (Papineni et al., 2002).

The shortcomings of these methods have been widely criticised and 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 neither correlate with human judgment on adequacy(whether the hypothesis sentence adequately captures the meaning of the reference sentence) nor on fluency(the quality of language in a sentence). Sulem et al (Sulem et al., 2018) examine BLEU – in the context of text simplification – on grammaticality, meaning preservation and simplicity, and report a very low, and in some cases, negative correlation with human judgment.

Considering these results it is a natural step to pursue new avenues for natural language evaluation and with the advent of deep learning using neural networks for this task is a promising step forward.

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.

Both of these methods report stronger correlations with human judgment and better results when compared to BLEU and ROUGE. While they are using word embeddings (Mikolov et al., 2013) to transfer their sentence in a continuous space they are still using distance metrics to evaluate that sentence. While BLEND (Ma et al., 2017) uses an SVM to combine different existing evaluation metrics.

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 evaluators success can be determined.

2.4 GLUE Benchmark

The GLUE Benchmark is a tool for evaluating and analyzing the performance of models across a diverse range of existing NLU tasks (Wang et al., 2018). The recent introduction of this benchmark has catalyzed the development of architectures scoring well on a wide variety of tasks and encouraged the NLP community to move away from specialized models doing well on a single task to models performing well across benchmarks. The variety of tasks introduced in the GLUE Benchmark are linguistic acceptability, sentiment analysis, semantic similarity, question answering, logical inference and reading comprehension. To be assessed according to that benchmark, models such as Transformers are usually pre-trained on a large corpus in an unsupervised manner and fine-tuned on a dataset used for the specific task of the benchmark.

3 Challenges with BLEU and ROUGE

In this part, we will discuss the limitations of BLEU and ROUGE. There are simple ways to attack these n-gram based metrics like adding a single word negation or changing all possible words with synonyms. Although these are theoretically plausible scenarios we also wanted to analyze which cases forced these metrics to fail in real life.

We took 100 examples from the STS-B dataset (Cer et al., 2017) where the absolute difference between the BLEU/ROUGE score and normalized label was the biggest. This does not necessarily capture all failure cases of BLEU/ROUGE but a variety of failure cases can be observed. Through this analysis, we see that there are systematically recurring real life examples, just like in our theoretical examples, where BLEU and ROUGE are failing to assess the level of similarity between two sentences.

We also observe that though some of the shortcomings of unigram metrics are mitigated through higher order n-grams, they open the door for different problems. Some of the most common failure cases that we have encountered are listed in table 1.

3.1 Idioms and adding details

One quite common problem we have is when idioms are used or extra examples/details are given in one of the sentences. These types of errors are especially common in more natural conversations. These types of errors also made 25 % of our analysis. Here we characteristically see humans giving high scores to these sentences because they are aware of which part holds the core meaning of the sentence where BLEU/ROUGE lack this ability.

One example from the dataset is “You should take this animal to a vet right away.” and “As covered in the other answers, your only option is to see a vet in order to have surgery done.” while the true score for this sentence pair is 3.6 out of 5 our BLEU score is 0.46 out of 5.

3.2 Changing words

Another of the most common error cases of BLEU and ROUGE is where one or a few important words of a sentence is changed while the rest of the structure is kept the same. There are many examples to this case and the common thing is we see the sentence structure preserved but words changed to alter the meaning remarkably. One example from STS-B is “a man is speaking.” and “a man is spitting.” while human judges give these two sentences a similarity score of 0.64 out of 5 our BLEU-1 score is a 3.75 out of 5.

3.3 General paraphrase

We also frequently see BLEU and ROUGE failing with general paraphrases which both include the above mentioned changes but also extend to words being replaced by synonyms, 12 % of cases, reordering of sub-sentences, around 10%, using different tenses or even simple spelling errors or differences, around 15%. These are not adversarial methods that are formulated to exploit any evaluation metric but direct results of the very nature of language.

While the higher order n-grams are supposed to preserve the intelligibility of the sentence and not reward a model that outputs words in a random order they also punish valid reorderings of sub sentences or words. In these smoothed methods changing a word with a synonym will also result in a much higher penalty. This higher penalty is supposed to work as a layer of protection for when only a single word is changed that shift the meaning of the sentence completely but this also results in a low similarity score if the meaning of the sentence stays exactly the same. Since this metric is oblivious to the meaning of the word that replaces the former let alone having a sense of context for this word.

BLEU and ROUGE are methods that are much more frequently under scoring sentence pairs than over scoring them hence similar to Reiter (Reiter, 2018) we conclude that BLEU/ROUGE can be fruitful in deciding whether a model is bad but not whether if it is good. While this analysis has been conducted in more detail the complete extend is beyond the purpose of this paper. The examples given above are fairly enough to show the main type of errors BLEU/ROUGE are facing and why they are falling short in the very task they were designed to do. We also provide detailed examples of these failure cases in the appendix.

3.4 Experiments

3.4.1 Some examples from STS-B sentences pairs

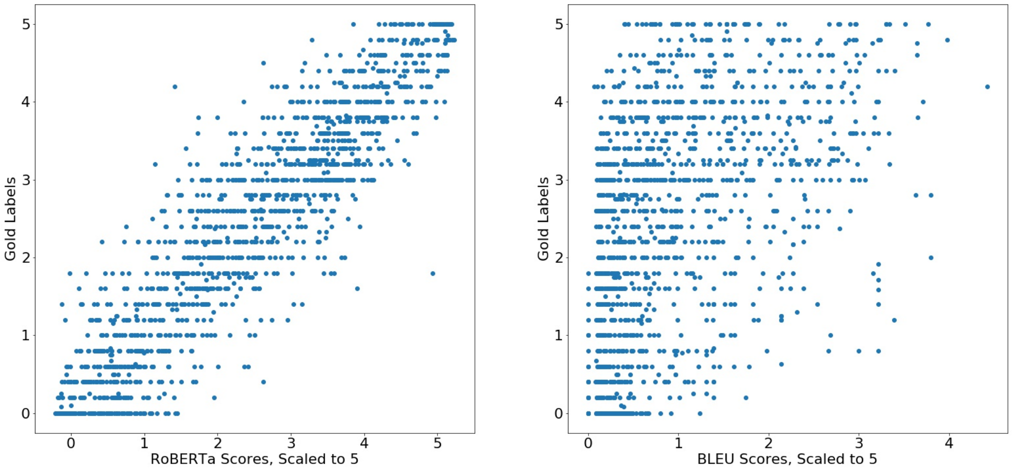
To illustrate our argument, we will give some examples from the dataset with their BLEU/ROUGE scores as well as a score generated from a RoBERTa model fine tuned on the STS-B dataset. Note that in this paper the BLEU* and ROUGE* scores are not between 0 and 1 but are scaled with 5 to be more understandable with the scale of the scoring metric used in the dataset. That is why we refer to them as BLEU* and ROUGE*

Table 1: BLEU*/ROUGE* and RoBERTa scores on sentence pair examples from STS-B

Sentence pair	**BLEU***	**ROUGE***	**RoBERTa**	**Label**
The last time the survey was conducted, in 1995, those numbers matched. In 1995, the last survey, those numbers were equal.	0.99	1.42	**4.65/5**	**5.00/5**
A band is performing on a stage. A band is playing onstage.	1.14	2.29	**3.85/5**	**5.00/5**
Two white dogs are swimming in the water. The birds are swimming in the water.	3.00	3.23	**1.19/5**	**0.80/5**
A man plays the piano. A man is playing a piano.	0.92	2.17	**5.00/5**	**5.00/5**
Pardon the brevity of this answer, but I would say ”named” is preferred within the context of your example. Named is preferred in your example, since you are formally giving a name to your method.	0.25	0.72	**3.73/5**	**4.40/5**

3.4.2 Semantic similarity experiments

In the above section we tried to give a sense of the mechanism of the failure of BLEU and ROUGE but in order to get a comprehensive view will also look at the general performance of these scoring methods.



(a)

(b)

Figure 1: Comparison of RoBERTa scores(a) and BLEU* scores(b) with labels from the STS-B dev set.

In figure 1 we can see that development set scores of RoBERTa in figure 1a and BLEU* in figurue 1b compared to the gold labels. While we will look at the correlation scores we can also see that the average error in BLEU and RoBERTa are remarkably different. This also clearly shows the difference in robustness between these two models and also emphasises that the failure cases we have given above are not exceptions but on the contrary are quite frequent.

Table 2: Correlation with human judgement of similarity on STS-B Benchmark development set

	ROUGE	BLEU	RoBERTa
Pearson correlation with human judgement	0.55	0.50	**0.92**

In table 2 we also look at the Pearson correlation between human judgment and different evaluators in the STS-B challenge development set.

4 Towards a robust, data-driven evaluation approach

4.1 Metric Scorecard

In our methodology to design new evaluation metrics for comparing reference summaries/translations to hypothesis ones, we established first-principles criteria on what a good evaluator should do. Key contributions of this section include separating the criteria of a good evaluator from the implementation of those criteria and moving away from uni-dimensional, human-designed evaluators towards multi-dimensional, model-based evaluators.

4.1.1 Motivation for multi-dimensional evaluators

Most metrics currently used have a stated goal of automatically determining the quality of a summary but fail to exactly mention the components of a high quality summary. This ambiguity contributes to these metrics not being robust as heuristic modifications made to ,for example, account for bi-grams and trigrams overlap fail to take into account other syntactic modifications such as changing a sentence from the active voice to the passive voice.

Even a data-driven evaluator such as a BERT semantic similarity score can be attacked if fed sentences which do not make grammatical sense but are assigned a higher than expected similarity score. For example, let's consider sentences s1: “A man is carrying a canoe with a dog” and s2 “A dog is carrying a man in a canoe with”. BERT-based similarity scores give s1 and s2 a similarity score of 5 whereas sentence s2 does not make grammatical sense.

To account for the flaws of any particular dimension and the evaluators only assessing for that dimension, we propose a modular, multi-dimensional criteria. The advantage of breaking down the definition of a high quality summary into modular criteria is that it simplifies the assessment of evaluation metrics and the models used to score each dimension can be recomposed and combined back into a single metric.

4.1.2 Initial criteria for robust evaluators

The first criteria of a good evaluation metric to compare reference translations/summaries against hypothesis summary is that it should be highly correlated with human judgement of semantic similarity. These automatic metrics were first designed because human evaluation of summary/translation is expensive and all state having a high correlation with human judgement of similarity as an objective (Lin, 2004) (Papineni et al., 2002) (Denkowski & Lavie, 2011).

As a complement to semantic similarity, having an evaluator able to distinguish linguistically similar sentences which are in logical contradiction will make it robust against flaws observed in previous section where sentences can have a high number of shared words but a few words which dramatically change their meaning as in the “NOT” attack case.

The third criteria comes from the observation that solely optimizing for BLEU/ROUGE has led to models able to achieve a high score without necessarily generating legible sentences (Paulus et al., 2017). As also shown above, solely optimizing for semantic similarity according to a model can also still lead to producing sentences which do not make grammatical sense.Having an evaluator able to penalize and identify hypothesis summaries which do not make grammatical sense will make evaluation robust against this flaw.

A fourth criteria of robust evaluators is that they should be difficult to game. That is, they should assign a high score to a reference summary against a hypothesis summary if, and only if these summaries have similar meaning and should assign a low score to a reference summary if, and only if, those summaries/translation have dissimilar meaning. Although obvious, we have empirically and theoretically shown that ROUGE and BLEU do not currently fit this criteria.

4.1.3 Using the metric scorecard in practice

We envision two primary use cases for the metric scorecard. It can serve as a benchmark to objectively assess and compare evaluation metrics such as ROUGE, BLEU, METEOR. It can also be used to leverage recent NLP models to design evaluation metrics as we explore in the next section.

4.2 Implementing model-based evaluators satisfying scorecard

After presenting the motivation for our multidimensional scorecard, we now proceed to show how it can be used to design summary evaluators. The scenario is that we have a reference summary s1 and a hypothesis summary s2. We want to know whether s2 is a high quality summary similar to s1. Until now, that definition was vague and mainly captured by the BLEU and ROUGE score as evaluators. In the previous section, we established 4 criteria for a good evaluators. 3 of those criteria can be used to assess how close s2 is to s1 given that the fourth criteria (robustness) is a property of the evaluator itself. The 3 criteria we have identified which are applicable to designing model-based evaluators include:

- Eval(s1,s2) should have one dimension assessing the semantic similarity of s1 and s2. This dimension should have high correlation with human judgement of similarity
- Eval(s1,s2) should have one dimension able to identify if s1 and s2 are in contradiction, unrelated or agreement.
- Eval(s2) should be able to identify is s2 makes grammatical sense and is not a gibberish sentence.

For these three criteria, we will show how we can use recent advances in language modelling to design evaluators which are either significantly outperforming BLEU/ROUGE on these dimensions or adding a criteria that BLEU and ROUGE could not previously control for when assessing hypothesis summaries.

4.2.1 Semantic Similarity

The first criteria of a good evaluator is that it should have a high correlation with human judgement of semantic similarity. To develop an evaluator corresponding to that dimension, we picked the best performing model on a semantic similarity task of the GLUE benchmark as of the writing of this paper.

Starting from the RoBERTa large pre-trained model (Liu et al., 2019), we finetune it to predict sentence similarity on the STS-B benchmark dataset (Cer et al., 2017). Given two sentences of text, s1 and s2, the system needs to compute how similar s1 and s2 are and returns a similarity score between 0 and 5. The dataset comprises naturally occurring pairs of sentences drawn from several domains and genres, annotated by crowdsourcing. The benchmark comprises 8628 sentence pairs with 5700 pairs in the training set, 1500 in the development set and 1379 in the test set.

4.2.2 Logical Equivalence

The second criteria of a good evaluator is that it should be able to accurately detect whether s1 is in contradiction with s2 or not. To develop an evaluator corresponding to that dimension, we picked the best performing model on a logical inference task of the GLUE benchmark as of the writing of this paper.

For logical inference, we start with a pretrained RoBERTa model and finetune it using the Multi-Genre Natural Language Inference Corpus (Nangia et al., 2017) . It is a crowdsourced collection of sentence pairs with textual entailment annotations. Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis, contradicts the hypothesis, or neither (neutral). The training set includes 393k sentence pairs, development set includes 20k and test set includes 20k. The accuracy of the pre-trained model on the development set is **0.9060**.

4.2.3 Sentence Intelligibility

The third criteria of a good evaluator is that it should be able to detect sentences which do not make grammatical sense. To develop an evaluator corresponding to that dimension, we explored two possibilities: directly using a state of the art language model or using an architecture fine tuned on a task such as linguistic acceptability.

For a language model based evaluator of sentence intelligibility, we can use an architecture such as GPT-2 (Radford et al., 2018) and use the perplexity score. The rationale for using the perplexity score of a large scale language models is that one way to frame a sentence which is not intelligible is that the order of words is surprising to a reader. Being surprising to a reader means having words not expected. Since language models trained to predict the next word maintain a probability distribution over an upcoming word given previous words, an unlikely word would surprise the language model.

This general insight is captured by the perplexity score assigned by a language model to a sentence. The current challenge is that perplexity scores alone can not inform whether a sentence makes grammatical sense. Assuming that the reference sentence has a correct syntax, one way of implementing the intelligibility criteria can be comparing the relative perplexity score of reference and hypothesis summaries.

For our fine tuned model, we start with a pretrained RoBERTa model and finetune it using the Corpus of Linguistic Acceptability (CoLA) . It consists of examples of expert English sentence acceptability judgments drawn from 22 books. Each example is a single string of English words annotated with whether it is grammatically possible sentence of English. The training set for CoLA has 10k sentences and the development set includes 1k sentences. The current model gets **67.8** percent accuracy on the test set showing that this is still an active area of research.

4.2.4 Rationale for Language Models

The overall rationale for using Transformer-based language models fine tuned for specific aspects of the scorecard is that recent work has shown that language models are unsupervised multi task learners (Radford et al., 2019) and can rediscover the classical NLP pipeline (Tenney et al., 2019). More specifically, in these models, there are localizable regions associated with distinct types of linguistic decisions suggesting that they can directly encode a range of syntactic and semantic information. Furthermore, the information is encoded in a natural progression POS tags processed earliest, followed by constituents, dependencies, semantic roles, and coreference. That is, it appears that basic syntactic information appears earlier in the network, while high-level semantic information appears at higher layers. By taking these pre-trained models and fine tuning them on a specific task, we make them pay attention to the correct level of abstraction corresponding to the scorecard.

4.2.5 Composing dimensions of an evaluator

In this section, we have shown how we could use state of the art language models to implement our evaluator. As long as the evaluators output a numeric score for each criteria, they can be aggregated through a weighted sum or other composition function into an aggregate score that can be used to summarize the quality of a hypothesis summary and comparing it to a reference summary.

5 Conclusion and future work

In this work, we have shown three main limitations of BLEU and ROUGE and proposed a path forward outlining why and how state of the art language models can be used as summary evaluators. While Transformers are currently the best performing architecture on the GLUE benchmark and have been used to implement each dimension of our evaluators, the framework and approach is independent current models and the family of evaluators it generates is meant to evolve with the field.

This work opens at least four follow up questions which can be explored in future work. Firstly, the proposed dimensions of the scorecard may be redundant and/or incomplete. Secondly, assuming a fixed scorecard, there might be better ways to robustly implement an evaluator assessing that criteria. For example, instead of using one model, ensembling approach may enable to overcome weaknesses of each fine tuned model. Thirdly, assessing published summarization models using that scorecard is a next step and would provide another analysis of the state of the art. Finally, finding out whether the model-based evaluators can be used to design an objective function to optimize for and generate good summaries is also an open research question. An interactive or reinforcement learning scenario in which a generator takes as input a text, produces a summary and get feedback by the evaluators would be interesting to explore. Open ended questions include both the convergence of such a training procedure and ,assuming convergence, whether the summaries produced by such a procedure will be deemed high quality by human evaluators.

REFERENCES

Chris Callison-Burch, Miles Osborne, and Philipp 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.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

Elizabeth Clark, Asli Celikyilmaz, and Noah 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*, pp. 2748–2760, 2019.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pp. 85–91. Association for Computational Linguistics, 2011.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. Blend: a novel combined mt metric based on direct assessment—casict-dcu submission to wmt17 metrics task. In *Proceedings of the second conference on machine translation*, pp. 598–603, 2017.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R Bowman. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. *arXiv preprint arXiv:1707.08172*, 2017.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

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

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

Elior Sulem, Omri Abend, and Ari Rappoport. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*, 2018.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

A Appendix

In the appendix we will discuss the failure cases of BLEU, ROUGE and RoBERTa in detail to provide a better understanding of how these models can fall short in language evaluation. This is important because a good metric scorecard has to represent the quality of an evaluator and this experiments are to show that our metrics cover many of the failure cases and can assess them without the burden of manually evaluating the outputs of every evaluator.

We will start by taking examples from the BLEU and ROUGE dataset. As in the paper the BLEU scores used are always a uniform average up to 4-grams and the ROUGE score is the average of ROUGE-1 and ROUGE-2. Both scores are scaled up to 5 to increase the interpretebility of the scores given that the labels in the similarty dataset are between 0 and 5.

Table 3: BLEU*/ROUGE* and RoBERTa scores on sentence pair examples from STS-B

Id	**Sentence pair**	**BLEU***	**ROUGE***	**RoBERTa**	**Label**
1	The company claims it's the largest single Apple VAR Xserve sale to date. The company claimed it is the largest sale of Xserves by an Apple retailer.	0.44	1.62	**4.36/5**	**5.00/5**
2	A woman puts flour on a piece of meat. A woman is putting flour onto some meat.	0.63	1.78	**5.07/5**	**5.00/5**
3	He later learned that the incident was caused by the Concorde's sonic boom. He later found out the alarming incident had been caused by Concorde's powerful sonic boom.	0.74	2.58	**4.96/5**	**5.00/5**
4	It indeed appears the Andromeda galaxy (M31) and The Milky Way (MW) are en route to a collision. In a few billion years, the Milky Way and Andromeda will collide.	0.20	1.09	**3.37/5**	**4.40/5**

5	You definitely do NOT want to be supporting your weight with your arms on the bike for normal riding. No, don't support your weight on your arms Your hands simply aren't really made for supporting all that weight.	0.28	1.13	**2.73/5**	**4.20/5**
6	7 detained for 'house sister' scandal China detains 7 for "house sister" scandal	0.33	1.52	**4.25/5**	**4.20/5**
7	A man plays the violin. A man is playing violin.	1.14	2.41	**5.12/5**	**5.00/5**
8	It is simply the number of balls bowled divided by the number of wickets taken. Bowling strike rate is defined for a bowler as the average number of balls bowled per wicket taken.	0.80	1.84	**3.83/5**	**4.40/5**
9	Police helicopter crashes into pub in Glasgow - several casualties Helicopter crashes into roof of Glasgow club	0.47	1.36	**3.58/5**	**4.00/5**
10	Oil falls in Asian trade Oil prices down in Asian trade	1.62	3.14	**4.89/5**	**5.00/5**
11	A skateboarder jumps off the stairs. A dog jumps off the stairs.	3.21	3.77	**1.09/5**	**0.80/5**
12	Wigan 3-2 Wolves: Match report, pictures & video highlights Arsenal 0-0 Chelsea: Match report, pictures & video highlights	3.39	3.26	**0.58/5**	**1.20/5**

In table 3 we see examples of many different error cases and in most sentences we also have more than one cause for the drastic difference between BLEU/ROUGE and the Label. For instance in rows 1 and 6 we see that the cause for the error is the reordering of sub-sentences, spelling/punctuation and newly introduced words that don't change the meaning but merely extend it. While BLEU and ROUGE are failing in these examples we see that the RoBERTa model scores similar to the label. In line 7 we can see that the RoBERTa model score is above 5. While this is because the model is designed as a regression model, as expected given the nature of the task.

In rows 2 and 7 we see that the main difference is the form or tense of the verb in a sentence. While using higher order n-grams enforces the goal of BLEU and ROUGE to not over score randomly shuffled sentences or remarkable meaning shift due to minimal changes in words such as in row 11. This also makes BLEU severely under score simple changes with synonyms or valid re-orderings as seen in the examples below. This characteristic of BLEU reinforces the point that BLEU and ROUGE are not useful in tracking the state of the art and comparing the best methods but are tools to weed out bad models fairly simply.

In rows 3 and 9 we see sentences that differ due to using descriptive phrases instead of a word or extending the sentence with more information. These types of errors changes are also caught with language models since we know they have the ability to hold the meaning of multiple words and incorporate them to reach a related word as in the famous example of king - men + woman = queen [cite word2vec].

In rows 4,8 and 5 we see general paraphrases with the same meaning represented in a generally different sentence. In all cases we see an drastic difference between BLEU/ROUGE and the label but these cases also unearth a specific characteristic of the neural evaluator. In 4 and 8 we see that the error of the RoBERTa model comparatively lower than row 5. While it is hard to determine the exact cause through only looking at these examples table 4 for the RoBERTa failure cases will make this case more compelling.

While language models have a general sense of the context in a given sentence they still lack a general knowledge of the world. Hence in the second sentence of row 5, because the words riding, bike, bicycle are missing the model has a hard time recognising that the second sentence is also about the same topic. To test this we added "while riding" or "on a bike" at the end of a sentence and the score immediately went up to 3.6/5 while barely changing the BLEU* and the ROUGE score. In row 4 and 8 however, the context of the sentence is defined explicitly with the key phrases. We see this bias affecting RoBERTa scoring in the examples below.

Table 4: BLEU*/ROUGE* and RoBERTa scores on sentence pair examples from STS-B

Id	**Sentence pair**	**BLEU***	**ROUGE***	**RoBERTa**	**Label**
1	It would be unusual for a snake to attack a stationary person. I'm no herpetologist, but in my experience, snakes are in the "you don't bug me, I won't bug you" category.	0.34	0.00	**1.4/5**	**4.20/5**
2	New UN peacekeeping chief named for Central African Republic UN takes over peacekeeping in Central African Republic	1.16	**2.39**	3.69/5	**2.00/5**
3	From Broadway comedies like "The Seven Year Itch" (1952), "Will Success Spoil Rock Hunter?" Playwright George Axelrod, who anticipated the sexual revolution with The Seven Year Itch and Will Success Spoil Rock Hunter?	**2.03**	1.31	3.16/5	**2.00/5**
4	a group of navy seals are singing A group of military personnel are playing in a brass quintet.	0.40	**1.45**	0.75/5	**2.40/5**

In the above examples we will find a two points that will helps us better understand the RoBERTa as a neural evaluator. First we see that the neural network sometimes lacks a sense of context that is not given in the sentence explicitly. While these language models are trained on a large corpus and capture a sense of the words and language we still see that their performance is not perfect. We see these examples in row 4 where the model cannot relate a navy seal as a military personnel. Or as in row 1 where the model cannot model an idiom.

The second and more critical place where we need further development is especially detecting whether the core argument/message in a sentence is the same beyond whether if they are talking about the same things. As in rows 2 and 3. We see the same landmark words and can clearly say that the sentences are talking about the same things but what a human can distinguish is that they are saying unrelated things. This is one of the key motivations in including the language inference task in the scorecard. Since detecting whether a pair of sentences are related on what level is a key part of detecting sentence similarity.

One last thing we will mention is that while RoBERTa and BLEU/ROUGE have different error cases their performance on these error cases is also remarkably different in favor of the former. Table 5 shows the mean error of BLEU* and the RoBERTa model on each others top 500, which is one third of the development set, error cases.

Table 5: Average error of BLEU* and RoBERTa in the their low scoring sets. With rows corresponding to which models failure cases and the columns to which model is used to score

	BLEU*	RoBERTa
BLEU*	2.93/5	**0.47/5**
RoBERTa	1.68/5	**0.89/5**

We see in table 5 that in BLEU* has a remarkable error in both its failure cases and also the failure cases of RoBERTa while RoBERTa outperforms BLEU* in each category.

While neural evaluators have also room for improvement we can with confidence say that they are outperforming classical methods and with a methodical way of improving them can bolster progress of NLP research.