Pre-trained language models evaluating themselves - A comparative study

Anonymous ACL submission

Abstract

Evaluating generated text received new attention with the introduction of model-based metrics in recent years. These new metrics have a higher correlation with human judgments and seemingly overcome many issues of previous n-gram based metrics from the symbolic age. In this work, we examine the recently introduced metrics BERTScore, BLEURT, NUBIA, MoverScore, and Mark-Evaluate (Petersen). We examined their sensitivity to different types of semantic deterioration (part of speech drop and negation), word order perturbations, word drop, and the common problem of repetition. No metric showed appropriate behaviour for negation, and further no metric was overall sensitive to the other issues mentioned above.

1 Introduction

Alongside with the current developments in Natural Language Generation (NLG), evaluating the quality of artificially generated text is an equally important (and ever harder) task in the field. N-gram based metrics, like BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004), come with severe drawbacks (Belz and Reiter, 2006; Reiter and Belz, 2009) and given the the increasing versatility of modern NLG systems, they are assumed to struggle even more (Zhang et al., 2020; Sellam et al., 2020). Architectures based on the Transformer (Vaswani et al., 2017), like BERT (Devlin et al., 2019) or the complete GPT series (Radford et al., 2018, 2019; Brown et al., 2020), have increased the quality of artificially generated text to an extent that even humans tend to struggle distinguishing natural from artificial texts (Clark et al., 2021). Based on these models, new metrics have been introduced, such as BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), NUBIA (Kane et al., 2020), MoverScore (Zhao et al., 2019), or Mark-Evaluate (Mordido and Meinel, 2020), claiming to increase correlation with human judgment. We examine the latter introduced metrics using synthetic data. The examination will include several practical problems commonly observed in NLG systems. The code of our experiments is publicly available on GitHub[1]

2 Related work

Caglayan et al. (2020) compared different metrics, including BERTScore regarding their sensitivity to specific impairments. Their experiment (related, but not similar to ours) indicated that BERTScore is more sensitive to the semantic integrity than n-gram based metrics. Another analysis by Kaster et al. (2021) provides an evaluation of model-based metrics based on linguistic properties of their input. They showed that even model-based metrics tend to behave differently regarding specific modifications to their input. Some metrics showed a higher sensitivity to semantics, while others showed higher sensitivity to syntactic issues. Eventually, ensembling methods were proposed to combine the strengths of metrics. Based on the CheckList library (Ribeiro et al., 2020), Sai et al. (2021) introduced a library for assessing NLG metrics via different perturbations to the input data. Multiple metrics, including model-based ones, were assessed, and neither of them did show a proper *overall* sensitivity to *all* modifications. The most severe issue was found in an overall insensitivity to negation. Contrary to our work, Sai et al. (2021) did not examine different degrees of perturbations. Sai et al. (2021) further underline the criticism of evaluating metrics according to their correlation with human judgments, which was already criticized in an in-depth analysis by Mathur et al. (2020) about applying correlation as an evaluation measure.

[1] See appended zip-file.

3 Materials and Methods

Additionally to describing the respective metric, an exact specification of the setup and model-specific details are reported in Appendix A.

BERTScore is a cosine-similarity based metric for which the input is encoded using RoBERTa embeddings (Liu et al., 2019). Recall and Precision are computed by summing over tokens and computing maximum similarity to each token from the other sentence. The result is averaged by the sentence length. For Precision, the sentence summed over is the reference sentence, and vice versa for Recall. F1 measure is the harmonic mean of the former two. Furthermore, inverse-document-frequency (idf) weighting can be applied to each maximal similarity in reference and precision, which is computed from the reference corpus.

MoverScore (MS) is based on the Word Mover's Distance (Kusner et al., 2015), an instance of Earth Mover's Distance (Rubner et al., 2000). It computes the minimal transportation cost necessary to transform one sentence into the other based on the distance between n-gram representations, additionally considering relative idf-weights. Representations are extracted from the last five layers of a DistilBERT model (Sanh et al., 2020).

Mark-Evaluate Petersen (ME-P, Mordido and Meinel, 2020) utilizes population estimators (Ricker, 1975) to score the quality of candidate-reference pairs. Since the population size is known prior to the estimate, the capture mechanism is based on whether a vector is inside the k-nearest-neighborhood of the opposite embedding set. The assumption that each sample is uniformly likely to be captured is intentionally violated. The deviation between known and estimated population size is computed to obtain the final score of the metric.

BLEURT (Sellam et al., 2020), in contrast to previous models, is a BERT model (RemBERT , Chung et al., 2020) specifically trained for evaluation. For adapting the model to the evaluation task, an additional training step is introduced in which artificially altered sentences are fed to the model alongside with the original ones to augment the evaluation process. Modification include dropping words from sentences, back-translating them or replacing random words with BERT predictions. A quality score can be computed based on different signals stemming from these alterations. These signals include metrics like BLEU, BERTScore and ROUGE, back-translation likelihood, a binary back-translation flag as well as entailment-flags. Further, the model is also fine-tuned on human ratings.

NUBIA (NeUral Based InterchangeAbility, Kane et al., 2020) is an ensemble metric consisting of three transformer-based models focussing on different aspects of the assessment: A pre-trained RoBERTa model, finetuned on STS-B (Cer et al., 2017), another pre-trained RoBERTa model, finetuned on MNLI (Williams et al., 2017), and a pre-trained GPT-2 model (Radford et al., 2019). The results are combined in an aggregator module and subsequently calibrated to fit in $[0, 1]$.

4 Experiments

For all our experiments we used the CNN/Daily Mail data set (Hermann et al., 2015) from `huggingface.datasets` as a reference corpus. Since it represents a corpus of high-quality news articles, it is ideally suited to use the scores of its original sentences as an upper bound for the evaluated metrics. We randomly sampled 2000 texts from this corpus for all of the models, except for NUBIA and ME-P.[2] Resulting scores from artificial impairments of different degrees can subsequently be compared to this upper bound. The modifications include the following different commonly observed flaws in NLG systems and the underlying language models:

Word Drop A random drop of words mimics general quality deterioration. The larger the intensity, the larger the drop probability gets. At the highest level, only a few tokens are left. This approach was inspired by Mordido and Meinel (2020) and Semeniuta et al. (2019).

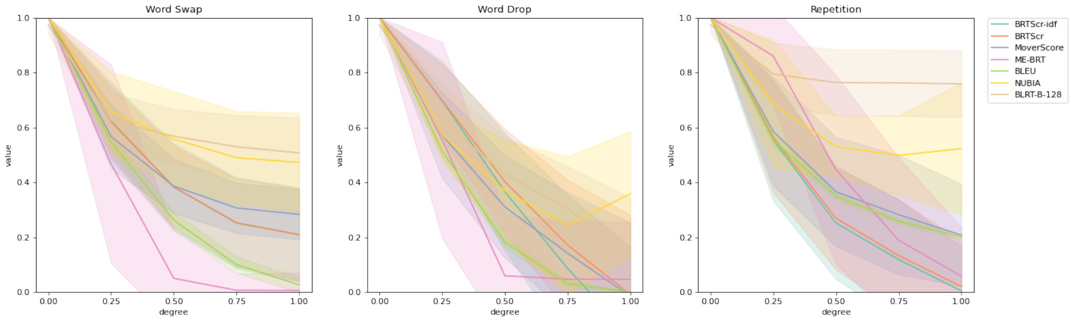
Word Swap Random word pairs are chosen and swapped. The higher the intensity, the more random the sequence of tokens becomes. Similar to word drop, this task was inspired by Mordido and Meinel (2020) and Semeniuta et al. (2019).

Repetition As shown by Fu et al. (2021), repetition remains a problem in text generated by NLG systems. A sequence at the end of the sentence

[2]NUBIA and ME-P are not optimized for use with GPUs, which is why we resorted to only using 50 of the 2000 texts.

[2]Examples for each of the different modifications are provided in Appendix B.

Figure 1: Development of the different metrics with increasing degrees of impairment



is chosen and repeatedly added to the sentence to mimic this issue. With increasing intensity, the chosen sequence is repeated more often and the overall sentence becomes longer.

Negation Sentences were negated to shift the semantics of the sentence into an entirely different direction. Negation is a minor sentence modification on a syntactic level, however, the sentence's semantics change entirely. For this modification, the CheckList library (Ribeiro et al., 2020) was utilized. This approach is analogous to the work of Sai et al. (2021).

POS-Drop Words with different part-of-speech (POS) tags were dropped to examine how the metrics behave, since some tokens are assumed to influence the degradation of overall semantic integrity more than others. SpaCy (Honnibal et al., 2020) and NLTK (Bird et al., 2009) were used to execute the different POS drops. As a baseline, the BLEU score is computed for each impairment which we then use for displaying the changes relative to BLEU (cf. Fig. 2).

5 Results

We expected to see a strict monotonous decrease for the impairments with increasing degree of severity. For Negation a sharp drop due to the deterioration of semantic meaning, while for POS-Drop the loss of rather unimportant POS (DET, ADJ) should intuitively not lead to more damage to the semantic integrity than the drop of important POS (NOUN, VERB).

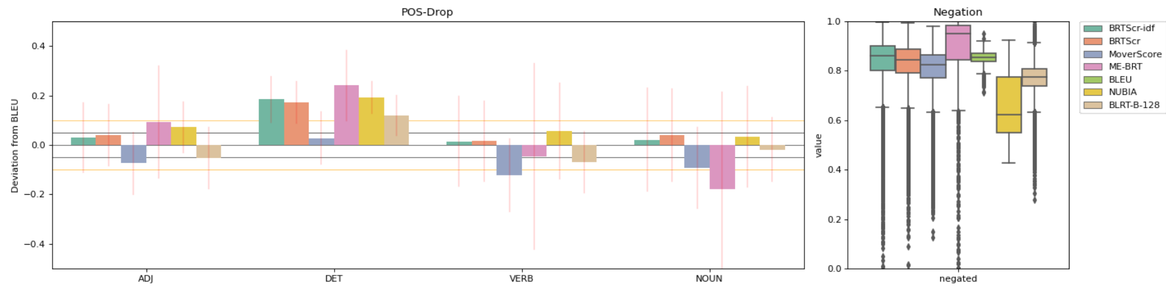
Results for continuous impairments (word drop, word swap and repetition) are displayed in Figure 1, while negation and POS drop are shown in Figure 2. For each type of impairment, we will report the most striking observations.

Swapped Words While BLEU exhibits, as expected, a steady drop to almost zero, some metrics tend to report higher values even when all words are swapped and the order is essentially random. NUBIA and BLEURT both have minimum values above 0.4, while MoverScore and BERTScore yield values above 0.2 for the highest degree of impairment. In contrast to this behavior, ME Petersen is most sensitive to word order perturbation and shows a sharp decline. It already drops to 0.47 at the first level of word order perturbation and reports a score of 0.01 for the random permutation.

Dropped Words In this task, BLEU, MoverScore, BERTScore, and ME-Petersen drop continuously until they eventually all (nearly) reach zero. ME-Petersen again drops the fastest, similar to the Word Swap but stops at 0.05. A different behavior, however, can be observed for BLEURT and NUBIA, which again exhibit higher values compared to the rest. BLEURT eventually drops to 0.14, and NUBIA even increases from its lowest value at the third level of impairment of 0.24 to 0.36 at the last level.

Repetition A less uniform behavior is observed for the repetition impairment, where the values strongly diverge at the highest level. Both BERTScore metrics monotonically decrease until they eventually reach zero, ME-Petersen also finally drops to a value near zero (0.06). However, it does not monotonically decrease, but drops sharply after the first level. BLEU and MoverScore both monotonically decrease strictly but end up way above zero at around 0.2. BLEURT and NUBIA behave entirely different, such that BLEURT seems to converge to 0.76 from the second level

Figure 2: Average Deviations (incl. Standard deviations) for all metrics relative to BLEU (for POS-Drop) and Boxplots for the impact of Negation on all metrics.



onward and does not show proper sensitivity to this issue, while NUBIA again increases after the third level from 0.5 to 0.52.

POS-Drop The most exceptional deviation from BLEU is observed in the removal of determiners. Most metrics (BERTScore, ME-P, BLEURT, and NUBIA) deviate positively from the reference, implying that the loss of determiner is less critical for the score, as expected. Adjectives, nouns, and verbs did affect metrics in different directions. Furthermore, BERTScore consistently reported higher values than BLEU.

Negation Since negation is a severe impairment to semantics, a significant drop in reported values was expected. However, the lowest reported score was observed in NUBIA, which dropped to an average of 0.65. BLEURT scores the second-lowest at an average of 0.77. All other metrics report an average between 0.81 and 0.86, including BLEU.

6 Discussion

Regarding word order perturbation, repetition, and word drop, it was expected to see a strict monotonous decline in the reported scores, which was not met by a single metric in every task (Although ME-P came close to meeting the expectations). However, for every task, at least one metric dropped to a value of zero or close to zero. However, one crucial aspect here is the metric-dependent sensitivity to word order perturbations and repetition, where especially the behavior of NUBIA and BLEURT is alarming. A further investigation of why both architectures behave differently from other representation-only-based metrics is thus needed in the future.

Our POS-drop task showed that some tokens influence scores more than others. Notably, the removal of determiners, which was expected not to influence the semantic integrity, did not lower the scores of most metrics. However, the syntactic integrity is affected, which must be considered when interpreting respective metrics. Behavior like this was also shown in Kaster et al. (2021) and was indicated by Caglayan et al. (2020) regarding BERTScore. No uniform behavior in most metrics was seen for removing verbs, nouns, and adjectives. Nonetheless, for nouns and verbs, the tendency to report a higher score is lower, which indicates a stronger emphasis on semantic integrity. However, sensitivity to semantic integrity is bound by the underlying model's capabilities, as observed in our negation task. No metric reported a proper value for the deterioration of semantic integrity, which aligns with Sai et al. (2021). The work of Kassner and Schütze (2020) and Ettinger (2020) already examined BERT regarding its understanding of negation, and they showed a general lack of understanding of the concept of negation.

7 Conclusion & Future work

Our results additionally underline that model-based metrics should be used with caution. The most severe drawback is the lack of sensitivity to negation, for which no metric reported a proper value. Hence further research in natural language understanding is necessary to overcome this issue. Furthermore, state-of-the-art metrics like BLEURT and NUBIA lacked sensitivity to repetition, which is a severe issue in NLG. Although many metrics deviated from the expected behavior, some others did not. Thus, we endorse the proposal of Kaster et al. (2021) to ensemble metrics and validate against the perturbation checklist package Sai et al. (2021).

References

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale.

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

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. Nubia: Neural based interchangeability assessor for text generation.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of bert-based evaluation metrics by disentangling along linguistic factors.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Gonçalo Mordido and Christoph Meinel. 2020. Mark-evaluate: Assessing language generation using population estimation methods. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1963–1977, Barcelona, Spain (Online). International Committee on 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 the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Association for Computational Linguistics (ACL)*.

William Edwin Ricker. 1975. Computation and interpretation of biological statistics of fish populations. *Bull. Fish. Res. Bd. Can.*, 191:1–382.

Yossi Rubner, Carlo Tomasi, and Leonidas Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.

Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. 2019. On accurate evaluation of gans for language generation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance.

Appendix

A Technical Setup

Metric	Underlying Model	Remarks
BERTScore (+ idf)	`microsoft/deberta-xlarge-mnli`	rescaled, hug_trns = 4.14.1, vers. = 0.3.11
BLEURT	`BLEURT-20`	finetuned RemBERT
Mark-Evaluate	`BERT-Base-MNLI`$^\heartsuit$	k = 1 (kNN)
MoverScore	`distilbert-base-uncased`$^\diamondsuit$	n = 1 (n-gram)
	`roberta-sts`	
NUBIA	`roberta-mnli`	
	`gpt-2`	sequences are clipped to max 1024 tokens

$^\heartsuit$ Available on GitHub
$^\diamondsuit$ As recommended in the official implementation

B Perturbation Examples

	Output
Original	`He’s quick, he’s a very complete player and in great form.`
Negation	`He’s quick, he’s not a very complete player and in great form.`
Repetition	`He ’s quick, he ’s a very complete player and in great form and in great form and in great form and in great form and in great form and in great form and in great form and in great form and in great form and in great form and in great form and in great form and in great form and in great form and in great form and in great form and in great form.`
Word Swap	`very complete a, he ’s quick He ’s and player great in form.`
Word Drop	`, player.`
Part of Speech Drop (ADJ)	`He’s he’s a very player and in form.`