Pre-trained language models evaluating themselves - A comparative study

Philipp Koch[♣] **Matthias Aßenmacher**[♠] **Christian Heumann**[♠]

Department of Statistics
Ludwig-Maximilians-Universität
Ludwigstr. 33, D-80539 Munich, Germany

[♣]P.Koch@campus.lmu.de, [♠]{matthias,chris}@stat.uni-muenchen.de

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 investigate 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 none of them 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 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 to reproduce 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. In contrast to Sai et al. (2021), our work focuses on examining different degrees of perturbations and how metrics reflect these modifications towards maximal impairment. 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. Furthermore, our work does not focus on correlation but solely on the scores which

[1] https://github.com/LazerLambda/MetricsComparison

Proceedings of the Third Workshop on Insights from Negative Results in NLP, pages 180 - 187

the different metrics report when confronted with specific impairments to various degrees, how metrics behave in contrast to BLEU when a particular part of speech is dropped, and how these metrics react to negated sentences.

3 Materials and Methods

The metrics examined in this work are BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), NUBIA (Kane et al., 2020), Mark-Evaluate Petersen (ME-P) (Mordido and Meinel, 2020), and MoverScore (Zhao et al., 2019). As a baseline, the BLEU score is always computed as well. The examined metrics can be subdivided into model-based metrics and metrics as trained models. NUBIA and BLEURT are trained models for evaluating generated text, while the other metrics are computed using specific formulas incorporating language models. Detailed descriptions of the metrics are provided in Appendix A. Additionally to describing the respective metric, an exact specification of the setup and model-specific details are reported in Appendix B.

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. The data set is in English entirely, i.e. all our findings do not necessarily transfer to other languages. 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[3] include the following different commonly observed flaws in NLG systems and the underlying language models:

Word Swap Random word pairs are chosen and swapped. The higher the intensity, the more random the sequence of tokens becomes, such that the original sequence should not be recognizable anymore. This approach was inspired by Mordido and Meinel (2020) and Semeniuta et al. (2019).

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. Similar to word swap, 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 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. At the maximum degree, the sequence is repeated as many times as there are tokens in the reference sentence.

Negation Sentences were negated to change the semantics severely. A simple syntactic change of the sentence has the power to shift the semantics in an entirely different direction. The CheckList library's (Ribeiro et al., 2020) experimental[4] negation function was utilized to apply this change. Specifically, the root of the dependency grammar tree is negated. This task was also used in 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 when different kinds of words are removed. We assume for our experiment that some part-of-speech units like determiners have less influence over the semantic integrity than the removal of verbs, nouns, or adjectives. SpaCy (Honnibal et al., 2020) and NLTK (Bird et al., 2009) were used to execute the different POS drops. The semantic-invariant and n-gram-based BLEU score is computed for each impairment, which we then use for displaying the changes relative to modern metrics. (cf. Fig. 2).

5 Results

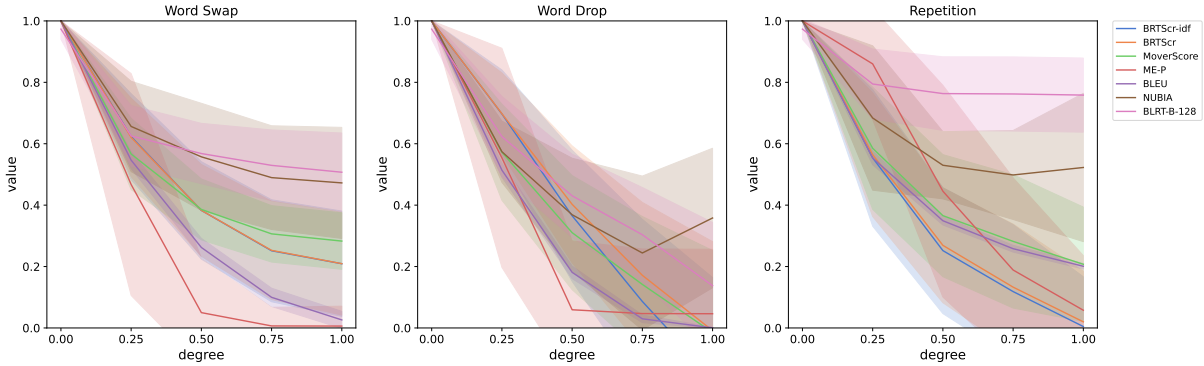
We expected to see a strict monotonous decrease for the impairments with increasing degree of severity. For Negation we expected a sharp drop due to the deterioration of semantic meaning. In the case of POS-Drop, the loss of rather unimportant POS (DET) should intuitively not lead to more damage to the semantic integrity than the drop of important POS (NOUN, VERB, ADJ), which is expected to be reported by the metrics as well. Furthermore,

[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.

[3] Examples for each of the different modifications are provided in Appendix C.

[4] See the respective notebook on GitHub.

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



the loss of different words should be reasonably comparable to BLEU.

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.

Word Swap 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.

Word Drop In this task, BLEU, MoverScore, BERTScore, and ME-P drop continuously until they eventually all (nearly) reach zero. ME-P 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-P 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 onward and does not show proper sensitivity to this issue, while NUBIA again increases after the third level from 0.5 to 0.52.

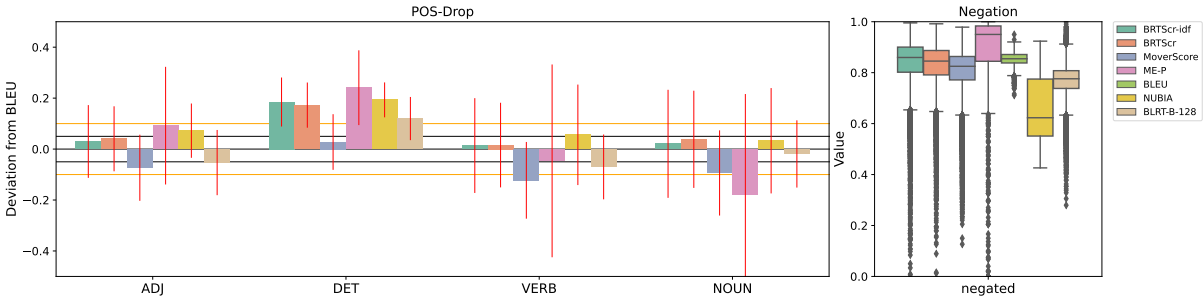
POS-Drop The most exceptional deviation from BLEU is observed in the removal of determiners (cf. Fig. 2). Most metrics (BERTScore, ME-P, BLEURT, and NUBIA) deviate positively from the reference, implying that the loss of determiner is less critical for the score, as expected. Adjectives, nouns, and verbs 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, at least one metric dropped to a value of zero or close to zero for every task. A crucial result is a metric-dependent sensitivity to word order perturbations and repetition. Especially for NUBIA and BLEURT, two trained metrics, the

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.



observed behavior 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 compared to BLEU. However, the syntactic integrity is affected, which must be considered when interpreting respective metrics. Semantic-focused 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. 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 severe semantic modification of negation, 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.

The most significant limitation of this work is the lack of expected ideal behavior when metrics are confronted with modified samples. It should be suspected that metrics show a higher drop in quality over more severe modifications, though it is unclear how humans would evaluate these specific cases. This issue is especially crucial in the task of negation since on the one hand side, it is not clear how severe the metrics are intended to reflect the impaired input, and on the other hand side it is also unclear how humans would rate negated sentences compared to the original sample. Consequently, the lack of human evaluation has to be considered when interpreting the results of this work. The same issue must be stated for POS-Drop tasks, in which human evaluation also becomes crucial. Further, it has to be taken into consideration that we use a feature described as experimental by its creators[5] for negating the sentences. Another arising issue, in this case, might be the rather long and detailed sentence structure of news article sentences, where the algorithm might be prone to negate only parts of the sentences. This issue might also arise for the POS-Drop case, since some POS units might occur more often in this data set than in other text.

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, since some showed strengths where others showed weaknesses, and validate against the perturbation checklist package Sai et al. (2021).

[5] See the respective notebook on GitHub.

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. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

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. *Transactions of the Association for Computational Linguistics*, 8:34–48.

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. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of BERT-based evaluation metrics by disentangling along linguistic factors. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8912–8925, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

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 Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Appendix

A Metrics

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. We use both a configuration without and with idf-weighting in our experiments.

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, fine-tuned on MNLI (Williams et al., 2018), 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 $[0, 1]$.

B Technical Setup

Table 1: Overview on the technical setup of the evaluated metrics.
♡ Available on GitHub
◇ As recommended in the official implementation

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`♡	k = 1 (kNN)
MoverScore	`distilbert-base-uncased`◇	n = 1 (n-gram)
	`roberta-sts`	
NUBIA	`roberta-mnli`	
	`gpt-2`	sequences are clipped to max 1024 tokens

C Perturbation Examples

Table 2: Examples of the different deteriorations. All other necessary details needed to reproduce our experiments can be found in the GitHub repository.

	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 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.`