NLG-METRICVERSE: An End-to-End Library for Evaluating Natural Language Generation

Giacomo Frisoni, Antonella Carbonaro, Gianluca Moro,
Andrea Zammarchi and **Marco Avagnano**
Department of Computer Science and Engineering (DISI)
University of Bologna, Via dell'Università 50, I-47522 Cesena, Italy
{giacomo.frisoni,antonella.carbonaro,gianluca.moro}@unibo.it
{andrea.zammarchi3, marco.avagnano}@studio.unibo.it

Abstract

Driven by deep learning breakthroughs, natural language generation (NLG) models have been at the center of steady progress in the last few years, with a ubiquitous task influence. However, since our ability to generate human-indistinguishable artificial text lags behind our capacity to assess it, it is paramount to develop and apply even better automatic evaluation metrics. To facilitate researchers to judge the effectiveness of their models broadly, we introduce NLG-METRICVERSE—an end-to-end open-source library for NLG evaluation based on Python. Our framework provides a living collection of NLG metrics in a unified and easy-to-use environment, supplying tools to efficiently apply, analyze, compare, and visualize them. This includes (i) the extensive support to heterogeneous automatic metrics with n-arity management, (ii) the meta-evaluation upon individual performance, metric-metric and metric-human correlations, (iii) graphical interpretations for helping humans better gain score intuitions, (iv) formal categorization and convenient documentation to accelerate metrics understanding. NLG-METRICVERSE aims to increase the comparability and replicability of NLG research, hopefully stimulating new contributions in the area. [1]

1 Introduction

Natural language generation (NLG) is a sub-field of natural language processing (NLP) concerned with automatically generating human-understandable text from input data, like prompts, tables, graphs, and images. Remarkably, the ability of a machine to produce text indistinguishable from that written by humans is a pre-requisite for Artificial General Intelligence (AGI)—the holy grail of AI. Recent advancements in deep learning have yielded tremendous improvements in the NLP sector, making NLG the object of fast-growing interest from the research community, as aptly demonstrated by GPT-3 (Brown et al., 2020). Pre-trained language models with transformer-based architectures (Kalyan et al., 2021) continue to push the envelope with unprecedented performance and encourage more and more applications. Indeed, today NLG includes a wide variety of tasks, such as machine translation, single/multi-document summarization, data-to-text, text-to-text, dialogue generation, free-form question answering, and image/video captioning (Gatt and Krahmer, 2018).

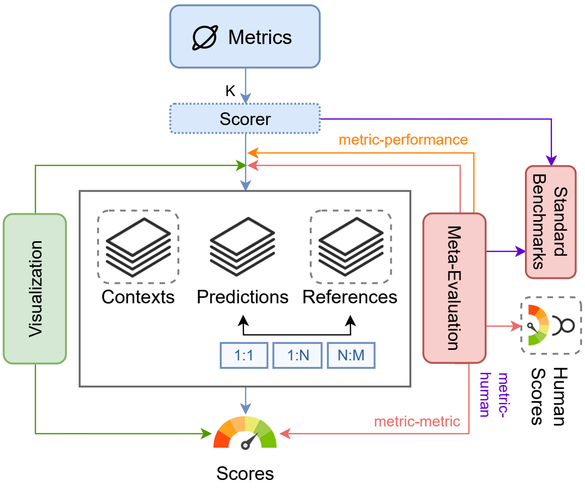


Figure 1: NLG-METRICVERSE operational representation. Dashed boxes denote optionality. A set of automatic metrics is selected to build a *Scorer* object, concurrently appliable to contexts, predictions, and references with arbitrary n-arity. A *Meta-Evaluation* module allows to inspect metrics' performance on the input data or standard benchmarks. Finally, a *Visualization* module can be applied to overcome opacity and understand metric-specific scoring processes.

As NLG models get better over time, accurately evaluating them is becoming an increasingly pressing priority for tracking progress in the area and convincingly recognizing state-of-the-art systems. However, the assessment of NLG model output is notoriously a challenging problem (Howcroft

[1] The code is publicly available at https://github.com/disi-unibo-nlp/nlg-metricverse

Proceedings of the 29th International Conference on Computational Linguistics, pages 3465–3479
October 12–17, 2022.

et al., 2020; Novikova et al., 2017). It involves the consideration of multiple intrinsic quality dimensions (e.g., informativeness, fluency, coherence, adequacy) and open-ended scenarios, where different plausible or equal-meaning responses may exist for the same user input. Human evaluation is typically regarded as the gold standard. Nevertheless, designing crowdsourcing experiments accompanied by elaborated guidelines is an expensive and high-latency process, which does not easily fit in a daily model development pipeline with the need for automatic benchmarking and tuning at scale. Furthermore, as NLG models improve, evaluators are asked to read longer passages of text conditioned on large amounts of context. In these cases, errors are often content-based (e.g., factual inaccuracies or context inconsistencies) rather than fluency-based, making superficial reads and non-expert annotators insufficient (Clark et al., 2021).

Given these issues, NLG researchers have settled for automatic evaluation metrics computing a holistic or dimension-specific score, an acceptable proxy for effectiveness and efficiency. Unfortunately, despite the rapid surge of machine-generated language, evaluation metrics have fallen behind, leaning on the conservative use of surface-level lexical similarities, which fail to cope with diversity and capture the text's underlying meaning. To overcome this severe bottleneck, the community has witnessed—in a relatively short time—a prolific, variegated, and original research production. New NLG metrics are constantly being proposed in top conferences, exhibiting one or more of the following characteristics: (i) use of contextualized word embeddings (Zhang et al., 2020), (ii) pre-training on massive unlabeled corpora (Sellam et al., 2020), (iii) fine-tuning on data annotated with human judgments (Kane et al., 2020), (iv) management of task-specific nuances (Dhingra et al., 2019; Wang et al., 2020).

Per contra, NLG metrics today are often designed and implemented from scratch with distinct environments, assumptions, properties, settings, benchmarks, and features. Such heterogeneity and disgregation make them difficult to compare or move to slightly different contexts. Concretely, the absence of a collective and continuously updated repository—well-documented and covering the entire NLG evaluation pipeline—discourages the use of modern solutions and slows down their understanding and practical application. Such barrier is highlighted also by the latest surveys (Sai et al., 2022). In the quest to fill this gap, we present NLG-METRICVERSE[2], an open-source (MIT licensed) end-to-end library for NLG evaluation, devised to provide a shared and collaborative codebase for fast application, analysis, comparison, visualization, and prototyping of automatic metrics.

The rest of the paper is organized as follows. First, we enumerate the design principles at the basis of NLG-METRICVERSE (§3), clarify the context, and summarize prior work related to this project (§2). Then, we describe the overarching NLG evaluation framework that constitutes the conceptual foundation for our contributions (§4). Next, we examine the main modules of the library: metrics, meta-evaluation, and visualization (§5). Lastly, we close the discussion and point out possible extensions (§7).

2 Background and Related Work

Early lexical NLG metrics, such as the BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), still appear to dominate the landscape, waiting for feasible, robust, and widely-adopted alternatives. Despite the high number of criticisms and studies proving their poor correlation with human judgment (Zhang et al., 2004), the popularity of first-generation metrics has not declined but expanded with the emergence of deep neural networks and new tasks. Simplicity, consistency, unsupervision, lightweight, and fast computation are the central basis of this success.

However, it has become increasingly clear that such adoption is often not prudent. Metrics measuring surface-level overlap are unsuitable for advanced evaluation, especially for modern text generation systems trained on mammoth data and with impressive paraphrasing capabilities (Mathur et al., 2020)—where ideal metrics should be sensitive to the underlying semantics. As a remedy, NLG researchers have started injecting learned/learnable components into their metrics, moving from a discrete space of word tokens to a continuous high-dimensional space of word vectors, thereby capturing distributional semantics. Over the years, many strong NLG evaluation metrics have been proposed, particularly transformer-based, like BLEURT (Sel-

[2] We coin the term "Metricverse" to denote the microcosm of automatic evaluation metrics powered by the overt ongoing rise of NLG models. According to this metaphor, we see metrics as planets belonging to galaxies and superclusters according to the taxonomy presented in Section 4.

lam et al., 2020), BERTScore (Zhang et al., 2020), and BARTScore (Yuan et al., 2021).

The trend towards the definition of model-based metrics and the resolution of task-specific needs have created a fertile ground for research. According to Sai et al. (2022), from 2002 (when BLEU was proposed) to 2014 (when Deep Learning became prevalent), there were only about 10 automatic NLG evaluation metrics in use; since 2015, at least 36 new metrics have appeared. On the other side, metrics are often scattered online, non-maintained, undocumented, implemented in various languages, inconsistent with the paper results. This not only hampers reproducibility but also inhibits scalability, as each research paper ends up creating its own implementation almost from scratch. Some libraries have already tried to make an integrated environment. To our best knowledge, NLGEval (Sharma et al., 2017), HugginFace Datasets (Lhoest et al., 2021), Evalaute[3], TorchMetrics (Detlefsen et al., 2022), and Jury (Cavusoglu et al., 2022) are the only resources available. However, none of them possess all the properties listed below: (i) large number of heterogeneous NLG metrics, (ii) concurrent computation of more metrics at once, (iii) support for multiple references and/or predictions, (iv) meta-evaluation, and (v) visualization. Table 1 summarizes the discrepancies between NLG-METRICVERSE and related work.

3 Design Principles

NLG-METRICVERSE has been designed with five main principles in mind, which, we argue, can help researchers and practitioners in a number of ways.

Comprehensiveness Given the impressive pace at which the field is growing, comprehensiveness is imperative, with the ultimate goal of providing a unique, smooth, and up-to-date access point to all the most relevant NLG evaluation metrics disseminated in different streams of literature. We also comprise organization and consistency across the library, with a coherent interaction between modules and sub-modules. This principle revolves around consolidating an all-in-one community-driven library, integrating ready-to-use n-gram- and embedding-based metrics—supervised and unsupervised, trained and untrained, reference- and statistics-based, task-specific and general-purpose, sentence- and document-level. From this synergy, we hope to spur the adoption of newly proposed contributions, unleashing their potential and concretizing the view of Sellam et al. (2020), according to which *"Machine Learning (ML) engineers should enrich their evaluation toolkits with more flexible, semantic-level metrics"*.

Ease-of-use The focus on simplicity is another key factor in fostering impact and usability, allowing users to write less code, reduce errors, and prototype faster. It is also meant to minimize the implementational burden and quickly move from papers to practical applications. We concentrate our efforts on designing an intuitive Application Programming Interface (API) accompanied by rich documentation with a curated list of executable notebooks and examples. This makes the software useful for both academia and industry.

Reproducibility Reproducibility is a core concept of utmost concern in ML and NLP, a prerequisite to trustworthiness. NLG evaluation exacerbates the problem even more, with well-known plagues like heavy undocumented preprocessing pipelines, non-transparent dataset selections, and concealed parameter settings (Post, 2018; Gao et al., 2021; Chen et al., 2022). A critical design objective of NLG-METRICVERSE is permitting experimental evaluation results to be seamlessly reproduced, promoting a fully detailed specification. In this way, users can simply integrate their original research into the shared codebase and fairly compare their solution with the existing literature. Besides serving for sound and consistent scientific research, reproducibility is a means to speed up the development of new metrics. When it comes to model-based metrics, transparency also applies to hardware setup, runtime measures, and CO2 impact.

Modularity In NLG-METRICVERSE, simplicity is sometimes bent in favor of modularity and reusability. This principle is essential for ensuring scalability and collaboratively bringing the codebase to maturity. An emphasis on module independence is maintained to guarantee the stand-alone usability of individual module functionalities and facilitate the learning of each library component.

Education One more principle is taking charge of an educational role. NLG-METRICVERSE is ideally suited to non-expert users, helping to sharpen their understanding. We believe that it is indispensable to democratize the field and gain greater

[3] https://github.com/huggingface/evaluate

	NLG-Metricverse	NLGEval	Datasets	Evaluate	TorchMetrics	Jury
#NLG-specific Metrics	38 + Datasets	8	22	22	13	19 + Datasets
More metrics at once	✓	×	×	✓	×	✓
Multiple refs/preds	✓	✓	×	×	×	✓
Meta-evaluation	✓	×	×	×	×	×
Visualization	✓	×	×	×	×	×

Table 1: Comparison of our library (v1.0.0) with existing NLG evaluation packages: NLGEval (v2.3.0), Datasets (v2.4.0), Evaluate (v0.2.2), TorchMetrics (v0.8.2), Jury (v2.2). "+ Datasets" stands for an automatic fallback towards HuggingFace Datasets in case of unsupported metrics (lower bound).

awareness of how metrics work. To unlock a teaching potential, our contributions want to include the release of standardized and content-rich metric cards, other than visualization tools conceived to aid unprecedented levels of score interpretation.

4 Framework

NLG-METRICVERSE is implemented as a Python library that provides a wrapper around a panoply of NLG evaluation metrics and complementary needs.

Regardless of the task, an NLG model generally produces one or more predictions (i.e., hypotheses, candidates) $p = p_1, \ldots, p_k$ conditioned on a given context or source $c = c_1, \ldots, c_p$. Then, one or multiple human-created references (i.e., ground-truths) $r = r_1, \ldots, r_l$ *may* be provided to assist the evaluation. In Table 2, we list sample contexts, predictions, and references for common NLG tasks to which NLG-METRICVERSE can be applied.

NLG Task	Context	Pred/Ref
Machine Translation	Source language sentence	Translation
Document Summarization	Document(s)	Summary
Data-to-Text	(Semi-)structured data, e.g., graphs, tables	Verbalization
Dialogue Generation	Conversation history	Response
Question Answering	Question (+ context)	Answer
Question Generation	Passage / Image / Knowledge Base	Question
Image/Video Captioning	Image / Video	Caption
Text Completion	Prompt	Continuation

Table 2: Popular NLG tasks settings.

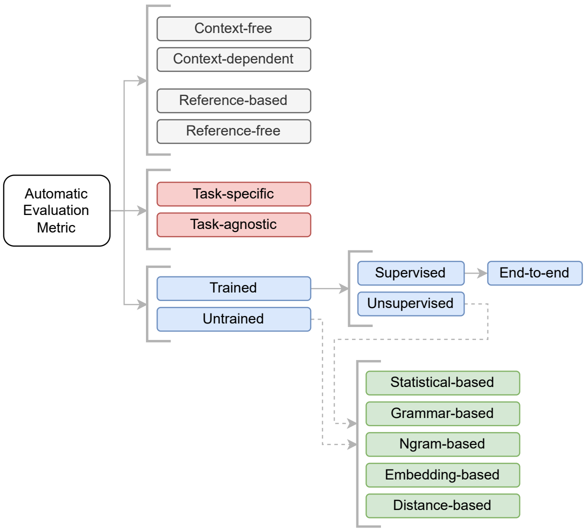


Figure 2: Taxonomy of automatic evaluation metrics. Different color nodes represent partially overlapping classification criteria (i.e., orthogonal categories).

Set these premises, NLG automatic evaluation metrics can be distinguished according to several overlapping criteria. To further dig into these distinctions, we lay out a taxonomy (Figure 2) serving as a foundation for experts and the broader public to build a shared overview of the possible solutions and their characteristics. Metrics can be broadly categorized based on the input format and data availability. *Context-free metrics* do not consider the context while judging the appropriateness of the prediction, typically being *task-agnostic* and adaptable to a wide variety of NLG tasks. On the flip side, *context-dependent* metrics take into account the context and are consequently *task-specific*. *Reference-based* metrics evaluate generated text with respect to one or a small set of reference text samples. *Reference-free* metrics do not rely on gold-standard references and are mainly statistics-based (e.g., full sequence distribution comparison). Further, they are suitable for an open-ended generation where there typically are several plausible continuations for each context, and creative generations are desirable; popular examples are Perplexity (Jelinek et al., 1977) and MAUVE (Pillutla et al., 2021). Finally, metrics can be classified according to their techniques. Metrics can have learnable components (*trained*) or not (*untrained*). In the first case, metrics can exploit human annotation data (*supervised*)—even with end-to-end architectures—or being human judgment-free (*unsupervised*). By end-to-end supervised metrics, we mean model-based NLG metrics trained on human-annotated data to directly output evaluation scores without additional techniques based on learned representations and placed outside the backpropagation process. They typically refer to solutions based on regression, ranking, and classification tasks (e.g., COMET (Rei et al., 2020), FactCC (Kryscinski et al., 2020), BLEURT (Sellam et al., 2020), NUBIA (Kane et al., 2020)). Untrained and unsupervised metrics use a fixed set of heuristics and input features, such as n-gram overlapping, edit distance, static or contextualized embeddings. In this context, grammar-based measures do not rely on ground-truth references and try to quantify aspects like readability (i.e., the ease with which a reader can understand a passage) and grammaticality. To provide a concrete example, BERTScore is a context-free, reference-based, trained and unsupervised metric.

5 Main Modules

NLG-METRICVERSE is organized into three main modules: Metrics (§5.1), Meta-Evaluation (§5.2), and Visualization (§5.3). The library is intended to be a continuous and collaborative project, extended as new solutions become available. In what follows, we describe the features provided at the current stage of development. Figure 1 shows the operational representation of the modules and their interplay within the framework detailed in §4. NLG-METRICVERSE is in turn built on top of open-source libraries, including Datasets (Lhoest et al., 2021), NumPy (van der Walt et al., 2011), SciPy (Virtanen et al., 2020), and Matplotlib (Hunter, 2007). Where possible, metrics are implemented using canonical repositories released by authors.

5.1 Metrics

To construct a full-scale NLG evaluation library, the selection methodology is crucial to collect metrics with desired properties. We concentrate on four factors. (i) *Diverse classes, supervision constraints, and evaluation tasks*, as defined in §4. NLG is a versatile field; the input/output scenarios and evaluation strategies can vary from case to case. Sometimes, the predicted text is short and accompanied by human target references; other times, diversity is preferred; still different times, the generation is open-ended, long, and without references. (ii) *Diverse application tasks*. Metrics can apply to multiple NLG evaluation tasks or manage task-specific quality needs. Hence, we include a broad spectrum of real-world tasks to boost the relevance of our library. (iii) *Eval dimension*. Evaluation can be done by assessing different quality perspectives. Most existing metrics cover a small subset of these axes. Still, some of them—particularly the trainable ones—can handle several dimensions by requiring to maximize correlation with each type of judgment separately (Rei et al., 2020) or not (Yuan et al., 2021). (iv) *Popularity*. We give priority to the metrics prominently used in NLG research. Currently, 34 metrics are supported (see §A.1 for details); more solutions are under development. We tried to cover a balanced mixture of metrics and paid importance not to overweight a specific class. Future contributions can easily be integrated into NLG-METRICVERSE. We ensure the integrity of each metric within the codebase through automated tests.

Input Format We design a unified metric input type, also handling n-arity for candidate and reference texts (Table 3)—a feature as vital as neglected by current systems. In fact, there may exist multiple equally good outputs for the given input, and comparing against one gold reference can be erroneous. An extensive set of out-of-the-box data loaders takes the responsibility of processing the raw data from files and directories.

Cardinality	Syntax
1:1	preds = $[p_1, \ldots, p_k]$, refs = $[r_1, \ldots, r_k]$

1:N	preds = $[p_1, \ldots, p_k]$
	refs = $[[r_{11}, \ldots, r_{1n}], \ldots, [r_{k1}, \ldots, r_{kn}]]$

N:M	preds = $[[p_{11}, \ldots, p_{1n}], \ldots, [p_{k1}, \ldots, p_{kn}]]$
	refs = $[[r_{11}, \ldots, r_{1m}], \ldots, [r_{k1}, \ldots, r_{km}]]$

Preds only	preds = $[p_1, \ldots, p_k]$

Table 3: Prediction-reference input formats.

Metrics Application Evaluating artificial text requires just two lines of code: (i) create a *Scorer* object with the desired metrics; (ii) apply the *Scorer* object to the input data. So, many metrics may be executed in one go. During step (ii), the proper strategy for computing each metric is automatically selected depending on the recognized input format. If a prediction needs to be compared against multiple references, the user is left with the possibility to specify the aggregation strategy of preference through the `reduce_fn` parameter. For example, `reduce_fn="max"` considers only the prediction-reference pair with the highest score for each dataset instance. Inherently, NLG-METRICVERSE allows all NumPy function names and custom aggregation functions as well. An asynchronous execution with a separate process for each metric can be specified to push efficiency and scalability (`run_concurrent`), bringing parallelism to the evaluation loop. Additionally, to contain the library size, we do not directly include all the packages required for running every supported metric, but we invite the user to install them if necessary. Figure 3 provides a practical example.

```
scorer =
 ↪  NLGMetricverse(metrics=["bertscore",
 ↪  "bartscore"], run_concurrent=True)
score = scorer(preds, refs) # reduce_fn
```

Figure 3: Definition and application of a *Scorer* object for the concurrent evaluation of multiple metrics.

By employing the `load_metric()` function for step (i), NLG-METRICVERSE falls back to the Datasets implementation in case of metrics not yet supported. Consequently, our library englobes at least any metrics that the Datasets package has. When defining the *Scorer*, a maximum degree of freedom is retained to allow the setting of metric-specific hyperparameters and different instantiations of the same metric (Figure 4). Further, since metrics generally involve several hyperparameters and results can deviate significantly for other choices, we accompany the output with a config report (hyperparams setting, hardware setup, etc.) for increasing comparability and replicability.

The *Scorer* application is meant to return a dictionary containing each metric's score(s), together with tracked performance metadata, including the computation time and CO2 emissions (measured with `codecarbon` (Schmidt et al., 2021)).

```
metrics = [
    load_metric("bleu",
     ↪  resulting_name="bleu_1",
     ↪  compute_kwargs={"max_order": 1}),
    load_metric("bleu",
     ↪  resulting_name="bleu_2",
     ↪  compute_kwargs={"max_order": 2}),
    load_metric("rouge")]
scorer = NLGMetricverse(metrics=metrics)
```

Figure 4: Definition and application of a *Scorer* object through the `load_metric()` function, encompassing two versions of BLEU with distinct hyperparameters.

Metric Documentation and Search NLP practitioners typically use automated metrics with a specific goal in mind, whether they are looking to answer a research question or develop a practical application system. To that end, they need to quickly identify which metric is most appropriate for the task at hand and understand how various attributes/properties might help with or, conversely, run contrary to their purpose. To let the user sift our NLG evaluation toolbox, we attach to each metric a set of structured tags (based on §4). Figure 5 exhibits APIs that allow users to list supported metrics and dig for those having preferred properties. We provide metric cards—inspired from aimed at evolving the Datasets ones—holding standardized[4] information about metric functioning, technical aspects, output bounds, etc. Since a metric's life continues beyond its initial release—from discovered weaknesses to newly found task adaptabilities, the metric card is conceived as a living document. The tags and metric cards are filled manually by the contributors who introduce the metrics to the library. The NLG-METRICVERSE community-driven nature and the GitHub-backend versioning provide an opportunity to keep the documentation up-to-date as further information comes to light.

```
NLGMetricverse.list_metrics()
# All
NLGMetricverse.filter_metrics(
 ↪  category=Categories.Embedding,
 ↪  appl_task= ApplTasks.DataToText)
# ["moverscore", "bleurt", "bartscore"]
NLGMetricverse.filter_metrics(
 ↪  trained=True, unsupervised=True,
 ↪  quality_dim=QualityDims.Factuality)
# ["bartscore"]
```

Figure 5: Taxonomy-guided metrics exploration.

Custom Metric NLG-METRICVERSE offers a flexible and uniform API for easily creating custom user-defined metrics. It only requires inheriting the `MetricForNLG` class (i.e., the common base class for each metric) and implementing the abstract functions linked to the possible input formats (Figure 6). We pursue the idea of enabling the user to create complex setups without superimposing constraints that may not suit future research.

```
class CustomMetric(MetricForNLG):
  def _compute_single_pred_single_ref(
     self, preds, refs, reduce_fn=None,
      ↪  **kwargs
  ): ...
  def _compute_single_pred_multi_ref ...
  def _compute_multi_pred_multi_ref ...
```

Figure 6: Custom metric implementation.

[4] https://bit.ly/metric-card-guideline

5.2 Meta-Evaluation

With the ever-growing number of proposed metrics, evaluating NLG evaluation has notoriously become a compelling exigency. The `meta_eval` module of NLG-METRICVERSE encompasses the most widely used methodologies for judging and comparing the effectiveness, reliability, and efficiency of automatic metrics. Few lines of code are sufficient to equitably assess a large number of published or prototype metrics on shared benchmarks.

Correlation Measures and Significance Tests Examining a set of NLG metrics usually presupposes the computation of different correlation measures on paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ depending on the goal and the relationship type between the two variables of interest X and Y. We support four standard correlation coefficients:

- *Pearson Correlation* (Freedman et al., 2007), measures the X-Y linear dependence;
- *Spearman Correlation* (Zar, 2005), measures the X-Y monotonic relationships (whether linear or not);
- *Kendall's* τ (Kendall, 1938) measures the X-Y ordinal association (ranking preservation);
- *DARR* (Ma et al., 2018), a robust variant of Kendall's τ to account for potential noise in Y through pairs filtering.

We refer the reader to Sai et al. (2022) for an in-depth discussion on their differences and selection criteria. In all cases, coefficients take values in $[-1, 1]$, from low to high agreement, with 0 denoting total independence. To compute the statistical significance of the quantified dependency strength, NLG-METRICVERSE considers the p-value of a hypothesis test examining the evidence against the null hypothesis that "population correlation coefficient equals 0". A smaller p-value means stronger evidence in favor of the alternative hypothesis, i.e., the population correlation is non-zero. The library also allows bootstrapping methods (Koehn, 2004) for rigorous pair-wise significance tests. Following previous works (Kilickaya et al., 2017; Novikova et al., 2017), we also incorporate the Williams' test (Williams, 1959) for evaluating the significance between two dependent correlations sharing one variable (i.e., X_1, X_2, and Y).

Metric-Human Correlation One of the primary goals of `meta-eval` is to analyze the extent to which different automatic evaluation metrics agree with human judgments (Figure 7). To do so, we provide tools for constructively computing metric-human correlations on popular benchmarks or custom user ground truths, where X and Y correspond to metric and human scores, respectively. As for benchmarking, we underline the urgency of standardized datasets containing `<context, prediction, reference, human scores>` tuples for multiple tasks, quality dimensions, and languages. The development of NLG evaluation metrics relies on their availability, both for training and evaluation purposes. Unfortunately, despite the evolving interest, there is still a scarcity of contributions in this direction. Currently, we use the annual public records from the WMT Metrics Shared Task (Bojar et al., 2017)—the largest collection of human ratings at the time of writing (i.e., human-annotated machine translation pairs).

```
metric_human_correlation(preds, refs,
↪   metrics=load_metric("rouge",
↪   compute_kwargs={"rouge_types":
↪   ["rougeL"]}),
↪   human_scores=Benchmarks.WMT17,
↪   corrs=[CorrelationMeasures.Pearson])
```

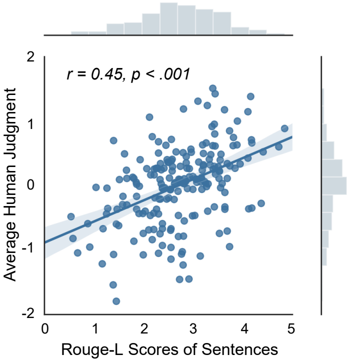
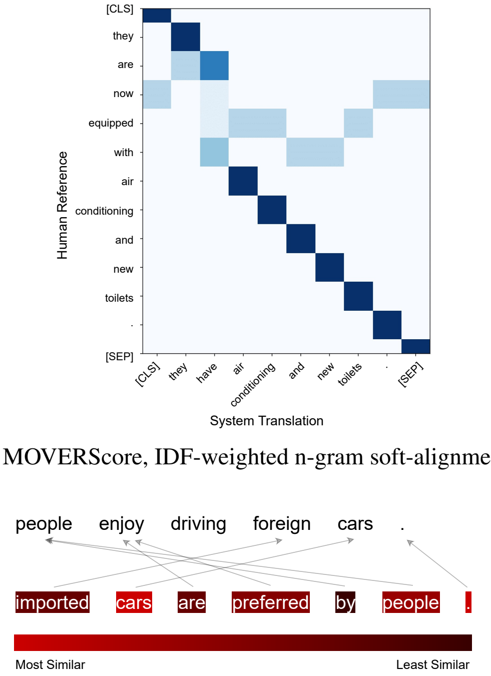


Figure 7: Segment-level metric-human correlation scatterplot. ROUGE vs. human scores on WMT17.

Metric-Metric Comparison On the trail of the most frequent evaluation setups used in literature, we supply functional features for checking out the behavior of many models side-by-side. In fact, metrics are best understood when compared to each other on common datasets. This comparison refers to performance aspects (e.g., computation time, CO2 impact for model-based metrics) and correlations (i.e., input-output similarities). Ultimately, NLG-METRICVERSE showcases the results with a set of meaningful charts intended to embolden scientific documentation (examples in Figure 9).

5.3 Visualization

In contrast to human evaluation, automatic metrics generally assign a single score to a given hypothesis, and it is often not clear which quality perspective this score captures or corresponds to; ergo, they are difficult to interpret (Sai et al., 2022). Score uninterpretability not only applies to contemporary model-based solutions but also to historical n-gram approaches (Zhang et al., 2004). More generally, visualization tools have become a cornerstone of explainability research in NLP. To increase the transparency of NLG evaluation metrics, we provide static and interactive visual tools for understanding *why* certain scores are produced. Visually inspecting internal mechanisms is particularly useful in instances when metrics disagree on. The interactive visualizations are built using web technologies manipulated through D3.js (Bostock et al., 2012). Supported ones include soft and hard alignments from MOVERScore and BERTScore (Figure 8).



(a) MOVERScore, IDF-weighted n-gram soft-alignment.

(b) BERTScore, Color-coded cosine similarity word matching.

Figure 8: Examples of plots for visual metric analysis.

6 Case Study: Graph-Augmented Biomedical Abstractive Summarization

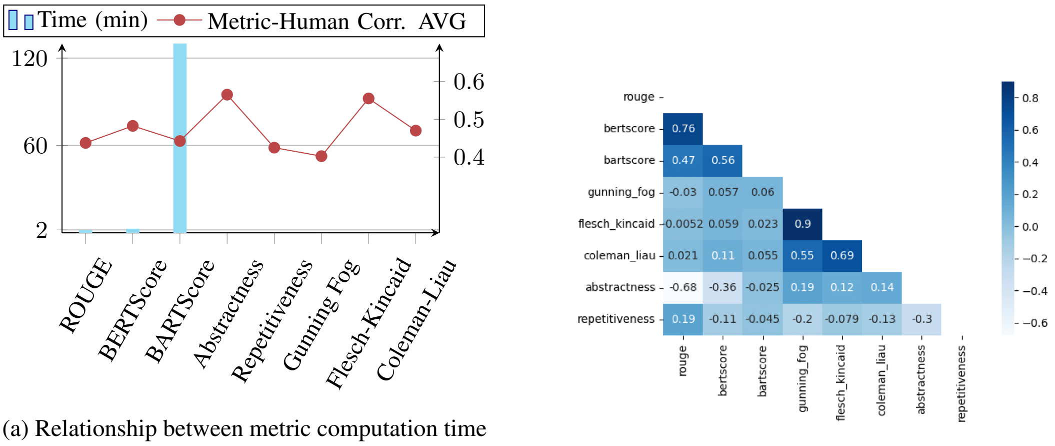
In this section, we use NLG-METRICVERSE to examine the summaries generated by a language model infused with semantic parsing graphs. Injecting explicit semantic structures—like events (Frisoni et al., 2021, 2022), abstract meaning representations (AMRs) (Banarescu et al., 2013), and corpus-level knowledge (Frisoni et al., 2020; Frisoni and Moro, 2020)—is a new trend followed by the NLP community to overcome lexical superficiality and draw a complementary path to architectural scaling, fundamental in low-resource settings (Moro and Ragazzi, 2022). Graph-augmented methods unlock a higher level of abstraction and more accurate emulation of human interpretation, rewriting, and paraphrasing. Faced with semantic-driven models, researchers must avoid being confined to traditional overlap-based metrics and monolithic quality dimensions, thus outlining a valuable testbed for our library.

6.1 Experimental Setup

We employ COGITOERGOSUMM (Frisoni. et al., 2022), a language model for biomedical single-document summarization, enhanced by AMRs and structured representations of factual evidence extracted from the source text. By employing the same hyperparameters proposed by the authors, we train and evaluate the neural network on CDSR (Guo et al., 2021)—a dataset designed for health literacy, where the training, validation, and test sets contain 5178, 500, and 999 samples, respectively. To quantitatively inspect model performance on the test set, we apply NLG-METRICVERSE for computing ROUGE-1/2/L (F1), BERTScore, BARTScore (Recall), Abstractness, and Repetitiveness. Additionally, since CDSR targets the accessibility of the biomedical literature, we calculate readability scores: Gunning Fog Index, Flesch-Kincaid Reading Ease, Coleman-Liau Index. See A.1 for details about metrics functioning, and A.2 for replicability. To better gauge summary quality and compare metrics' effectiveness, we conduct a human evaluation study. We randomly select 30 test set instances, and invite 3 expert annotators to score generated summaries in conformity with four independent perspectives, each measured on a Likert scale from 1 (worst) to 5 (best): (i) *informativeness*, i.e., conveying salient content; (ii) *factualness*, i.e., being faithful with respect to the article; (iii) *fluency*, i.e., being fluent, grammatical, and coherent; (iv) *succinctness*, i.e., non containing redundant and unnecessary information.

6.2 Results

Figure 9 reports human and automatic evaluation results, together with computation times, metric-metric, and metric-human correlations (Pearson).



(a) Relationship between metric computation time and average correlation with human judgment.

(b) Heatmap with pairwise metric correlations.

Human	Informativeness	Factualness	Fluency	Succinctness
	3.67	3.61	3.61	3.50
Auto	ROUGE-1/2/L	BERTScore	BARTScore	Abstractness
	0.49/0.19/0.25	0.87	-2.68	0.36
	Repetitiveness	Gunning Fog Index	Flesch-Kincaid	Coleman-Liau Index
	0.37	13.45	12.64	13.84

(c) Qualitative and quantitative evaluation scores.

Figure 9: Abstractive summarization analysis through NLG-METRICVERSE.

Human scores are averaged for each dimension; the mean Kendall coefficient among all evaluators' inter-rater agreement is 0.16. We observe that the abstractive and semantically-consistent nature of the model is not appreciable by the ROUGE scores alone. The highest correlations with human judgment are achieved by BERTScore, Abstractness, and Flesch-Kincaid—especially according to factualness and succinctness (see A.2). These results prove that the model tends to be more factual when it re-frames the target concept units, further testifying the inadequacy of overlap-based metrics. Notably, in contrast to other model-based metrics like BERTScore, BARTScore appears significantly slower (72× compared to ROUGE).

7 Conclusion

The NLG evaluation community demands efforts toward making research more transparent, reproducible, and open. Easy access to a wide variety of automatic metrics and related features holds a lot of potential. A central hub would democratize research, increase comparability, mitigate the computational/implementational burden, and hopefully steer innovation to more robust contributions. In fact, researchers would be able to evaluate their NLG systems at scale without being limited to very few metrics whose code is easily available. They would also be able to critically examine existing metrics, perform white-box attacks, or carefully craft adversarial examples.

With NLG-METRICVERSE, we take an important step towards a single, unified, coherent, end-to-end, and easily extendable framework for NLG evaluation. A solid reference point and shared resource for researchers and practitioners working in the area. Being a community-driven effort, we plan in both the near and medium terms to support more recent task-specific metrics, benchmarks, meta-evaluation techniques for robustness, and skew factor analyses. We also intend to include more document-level measures. We hope that this library may trigger a positive reinforcement loop within our community, nudging it to explore the metric universe.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR:

An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ondrej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 489–513. Association for Computational Linguistics.

Mike Bostock et al. 2012. D3. js-data-driven documents. *Project homepage at http://d3js. org*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Devrim Cavusoglu, Fatih Cagatay Akyon, Ulas Sert, and Cemil Cengiz. 2022. Jury: Comprehensive NLP Evaluation toolkit.

Yanran Chen, Jonas Belouadi, and Steffen Eger. 2022. Reproducibility issues for bert-based evaluation metrics. *CoRR*, abs/2204.00004.

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.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. 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.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Nicki Skafte Detlefsen, Jirí Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. 2022. Torchmetrics - measuring reproducibility in pytorch. *J. Open Source Softw.*, 7(69):4101.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

David Freedman, Robert Pisani, Roger Purves, and Ani Adhikari. 2007. Statistics (international student edition).

Giacomo Frisoni., Paolo Italiani., Francesco Boschi., and Gianluca Moro. 2022. Enhancing biomedical scientific reviews summarization with graph-based factual evidence extracted from papers. In *DATA*, pages 168–179. INSTICC, SciTePress.

Giacomo Frisoni and Gianluca Moro. 2020. Phenomena Explanation from Text: Unsupervised Learning of Interpretable and Statistically Significant Knowledge. In *DATA (Revised Selected Papers)*, volume 1446, pages 293–318. Springer.

Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2020. Learning Interpretable and Statistically Significant Knowledge from Unlabeled Corpora of Social Text Messages: A Novel Methodology of Descriptive Text Mining. In *DATA 2020 - Proc. 9th Int. Conf. Data Science, Technol. and Appl.*, pages 121–134. SciTePress.

Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2021. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. *IEEE Access*, 9:160721–160757.

Giacomo Frisoni, Gianluca Moro, Giulio Carlassare, and Antonella Carbonaro. 2022. Unsupervised event graph representation and similarity learning on biomedical literature. *Sensors*, 22(1):3.

Yang Gao, Steffen Eger, Wei Zhao, Piyawat Lertvittayakumjorn, and Marina Fomicheva, editors. 2021. *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics, Punta Cana, Dominican Republic.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Sebastian Gehrmann, Zachary Ziegler, and Alexander Rush. 2019. Generating abstractive summaries with finetuned language models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 516–522, Tokyo, Japan. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

R.; et al Gunning. 1952. Technique of clear writing.

Yinuo Guo and Junfeng Hu. 2019. Meteor++ 2.0: Adopt syntactic level paraphrase knowledge into machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 501–506, Florence, Italy. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *AAAI*, pages 160–168. AAAI Press.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

John D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.*, 9(3):90–95.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and S. Sangeetha. 2021. Ammus : A survey of transformer-based pretrained models in natural language processing. *ArXiv*, abs/2108.05542.

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.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain. Association for Computational Linguistics.

J. Peter Kincaid, Robert P. Fishburne, R L Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 175–184. Association for Computational Linguistics.

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.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on*

Machine Translation: Shared Task Papers, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

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.

Gianluca Moro and Luca Ragazzi. 2022. Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In *AAAI*, pages 11085–11093. AAAI Press.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004a. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Andrew Cameron Morris, Viktoria Maier, and Phil D. Green. 2004b. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*. ISCA.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Victor Schmidt, Kamal Goyal, Aditya Joshi, Boris Feld, Liam Conell, Nikolas Laskaris, Doug Blank, Jonathan Wilson, Sorelle Friedler, and Sasha Luccioni. 2021. CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing.

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.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. EED: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520, Florence, Italy. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Stéfan van der Walt, S. Chris Colbert, and Gaël Varoquaux. 2011. The numpy array: A structure for efficient numerical computation. *Comput. Sci. Eng.*, 13(2):22–30.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Evan James Williams. 1959. *Regression analysis*, volume 14. wiley.

Wen Xiao and Giuseppe Carenini. 2020. Systematically exploring redundancy reduction in summarizing long documents. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 516–528, Suzhou, China. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

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*. OpenReview.net.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

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.

A Appendix

A.1 Supported Metrics

Table 5 and Table 6 enumerates the metrics currently supported by NLG-METRICVERSE.

A.2 Case Study Replicability and Details

We used NLG-METRICVERSE on a workstation having one Nvidia Tesla T4 GPU with 16GB of dedicated memory, and an Intel® Xeon™ CPU @ 2.20GHz. Where applicable, we ran the metrics on GPU. For the sake of reproducibility, Table 4 lists all metrics' hyperparameters. Please note that ROUGE, BERTScore, Abstractness, and Repetitiveness bounds are in $[0, 1]$, BARTScore in $]-\infty, 0[$. Gunning Fog Index, Flesch Kincaid Reading Ease, and Colemain-Liau Index estimate the years of education generally required to understand a text document; lower scores indicate that the text is easier to read (U.S. college-level readability belongs to the range $[13-16]$).

Metric	Hyperparameters
ROUGE	rouge_types=["rouge1","rouge2","rougeL"], use_aggregator=True, use_stemmer=False, metric_to_select="fmeasure"
BERTScore	lang="en", idf=False, batch_size=64, nthreads=4, rescale_with_baseline=False, use_fast_tokenizer=False, return_average_scores=False
BARTScore	model_checkpoint="bartscore-large-cnn", batch_size=4, segment_scores=False
Abstractness	ngrams=1

Table 4: Hyperparameters initialization for metrics applied in the case study.

Metric	Technique	Property	Appl. Tasks	Trained	Unsupervised
Gunning Fog Index Gunning 1952	G	readability test for English writing: count of sentences, words, and complex words consisting of three or more syllables in the text	SUM	×	✓
Flesch-Kincaid Kincaid et al. 1975	G	the most widely used readability test for English writing; two versions (Flesch Reading-Ease and Flesch-Kincaid Grade Level)	SUM	×	✓
Coleman-Liau Index Coleman and Liau 1975	G	character-based readability test for English writing	SUM	×	✓
Accuracy Pedregosa et al. 2011	N	proportion of correct predictions among the total number of cases processed	MT	×	✓
Precision Pedregosa et al. 2011	N	fraction of correctly labeled positive examples out of all of the examples that were labeled as positive	MT	×	✓
Recall Pedregosa et al. 2011	N	fraction of positive examples correctly labeled by the model as positive	MT	×	✓
F1 Pedregosa et al. 2011	N	harmonic mean of the precision and recall	MT	×	✓
MER Morris et al. 2004a	N	% words incorrectly predicted and inserted (match error rate)	SR	×	✓
Abstractness Gehrmann et al. 2019	N	% novel n-grams in the predictions, compared to the references	SUM	×	✓
Repetitiveness Xiao and Carenini 2020	N	average number of n-grams with at least one repetition in the generated sequences	SUM	×	✓
Coverage Grusky et al. 2018	N	% summary words present in the source text	SUM	×	✓
Density Grusky et al. 2018	N	average length of extracted fragments which every word from the summary belongs to	SUM	×	✓
Compression Grusky et al. 2018	N	ratio between the length of the original text and the length of the generated abstract	SUM	×	✓
BLEU Papineni et al., 2002	N	n-gram precision	MT, IC, DG, QG, RG	×	✓
NIST Doddington 2002	N	n-gram precision w/ IDF-weighted n-grams	MT	×	✓
ORANGE (SentBLEU) Lin and Och 2004	N	n-gram precision w/ smoothing	MT	×	✓
ROUGE Lin, 2004	N	n-gram recall	MT	×	✓
WER Morris et al. 2004b	N	% of insert, delete, replace	MT, SR	×	✓
METEOR Banerjee and Lavie 2005	N	n-gram harmonic mean w/ paraphrase knowledge (e.g., stemming, synonyms) and penalty factor for fragmented matches	MT, IC, DG	×	✓
CIDEr Vedantam et al. 2015	N	cosine similarity between TF-IDF weighted n-grams	IC	×	✓
TER Snover et al. 2006	N	translation edit rate (i.e., WER + shift movement as extra editing step)	MT	×	✓
ChrF(++) Popović 2017	N	character-level precision and recall	MT, IC, SUM	×	✓
WMD Kusner et al. 2015	E, D	earth mover's distance on words	IC, SUM	×	✓
SMS Clark et al. 2019	E, D	earth mover's distance on sentences	IC, SR, SUM	×	✓
CharacTER Wang et al. 2016	N	character-level TER	MT	×	✓
SacreBLEU Post 2018	N	standardized BLEU	MT	×	✓
METEOR++ Guo and Hu 2019	N	METEOR w/ copy knowledge and syntactic-level paraphrase matching	MT	×	✓

Table 5: NLG-METRICVERSE supported metrics for the v1.0.0 release, in ascending order of publication. We use the following abbreviations for different techniques and features: G – Grammar-based, N – N-gram-based, D – Distance-based, E – Embedding-based, S – Statistics-based. For tasks, SUM – Summarization, MT – Machine Translation, SR – Speech Recognition, IC – Image Captioning, DG – Document or Story Generation, QG – Query Generation, RG – Dialogue Response Generation, D2T – Data-to-Text, TC – Text Completion; we only list the ones justified by the original paper or by the first NLG application.

Metric	Technique	Property	Appl. Tasks	Trained	Unsupervised
MOVERScore Zhao et al., 2019	E	IDF-weighted n-gram soft-alignment (WMD generalization) via contextualized embeddings; it computes the minimum cost of transforming the generated text to the reference text, taking into account Euclidean distance between vector representations of n-grams, as well as their document frequencies	MT, SUM, D2T, IC	✓ ELMo/BERT	✓
EED Stanchev et al. 2019	D	Levenshtein distance + jump operation	MT	×	✓
COMET Rei et al., 2020	E	multilingual-MT human judgment predictions through pre-trained cross-lingual encoders (word embeddings) + pooling layers (sentence embeddings) + feed-forward regressor or triplet margin loss depending on the judgment type (real-value or relative ranking)	MT	✓ XML-RoBERTa end-to-end	×
FactCC(X) Kryscinski et al. 2020	E	weakly-supervised document↔summary-sentence factual consistency evaluation based on BERT's [CLS] embedding	SUM	✓ BERT end-to-end	×
BLEURT Sellam et al., 2020	E	robust human score prediction based on fine-tuning a BERT model with an additional pre-training scheme characterized by millions of synthetic reference-candidate pairs and lexical-/semantic-level tasks combined through an aggregated loss	MT, D2T	✓ BERT end-to-end	×
NUBIA Kane et al. 2020	E	human score prediction with three modules: neural feature extractor on reference-hypothesis pairs (multiple pre-trained transformers capturing semantic similarity, logic entailment, sentence intelligibility) + aggregator (features→quality score mapping) + calibrator	MT, IC	✓ RoBERTa GPT-2 end-to-end	×
BERTScore Zhang et al., 2020	E	IDF-weighted n-gram hard-alignment via contextualized embeddings	MT, IC	✓ BERT	✓
BARTScore Yuan et al., 2021	E	multi-perspective evaluation as text generation via a pre-trained seq2seq model to measure how likely hypothesis and reference are paraphrased according to the probability of one giving the other	MT, SUM, D2T	✓ BART	✓
Perplexity Jelinek et al., 1977	E	how likely a model is to generate the input text sequence	SR	✓	✓
PRISM Thompson and Post 2020	E	sequence-to-sequence paraphraser to score MT system outputs conditioned on their respective human references	TC	✓ GPT-2 Grover	✓
MAUVE Pillutla et al. 2021	E, D	comparison measure for open-ended text generation w/ divergences in a quantized embedding space	TC	✓ GPT-2 Grover	✓

Table 6: Table 5 continuation.

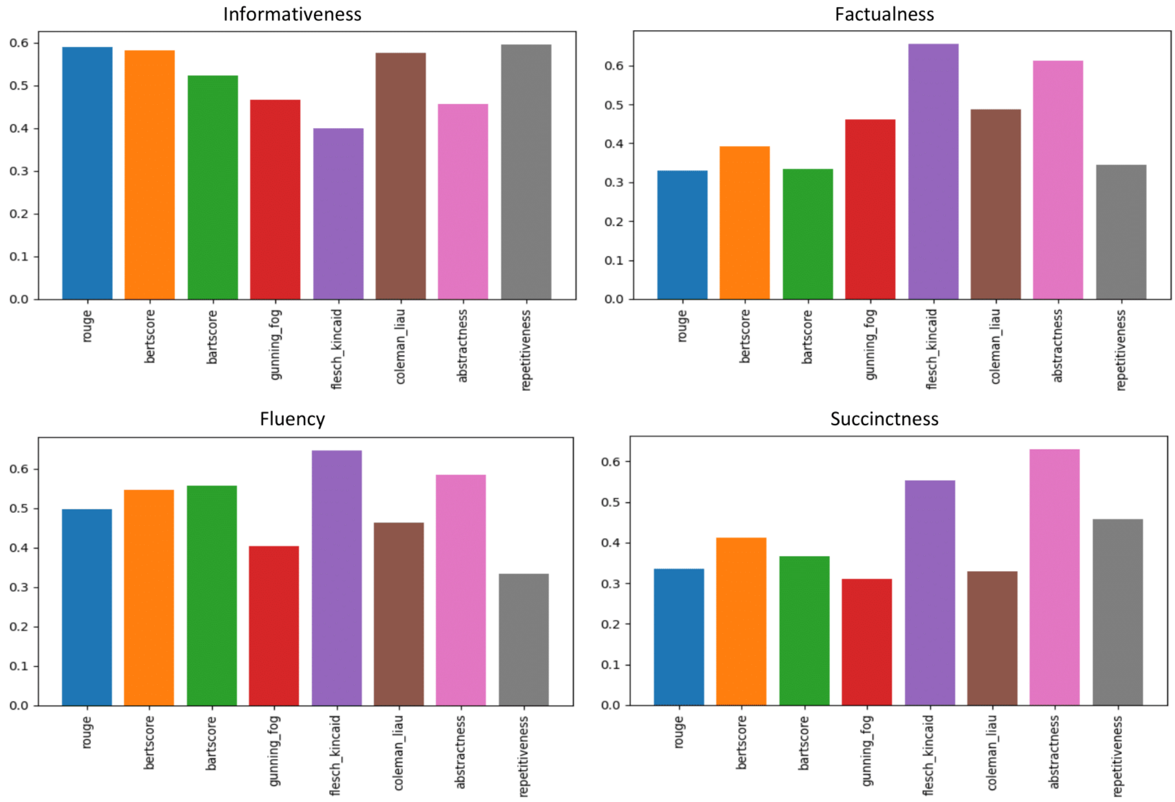


Figure 10: Pearson correlations between automatic metrics and human annotations for each quality dimension inspected in the case study, i.e., informativeness, factualness, fluency, succinctness.