

Reference-free Evaluation Metrics for Text Generation: A Survey

Takumi Ito
 Tohoku University
 Langsmith Inc.
 t-ito@tohoku.ac.jp

Kees van Deemter
 Utrecht University
 c.j.vandeemter@uu.nl

Jun Suzuki
 Tohoku University
 RIKEN
 jun.suzuki@tohoku.ac.jp

Abstract

A number of automatic evaluation metrics have been proposed for natural language generation systems. The most common approach to automatic evaluation is the use of a reference-based metric that compares the model’s output with gold-standard references written by humans. However, it is expensive to create such references, and for some tasks, such as response generation in dialogue, creating references is not a simple matter. Therefore, various **reference-free** metrics have been developed in recent years. In this survey, which intends to cover the full breadth of all NLG tasks, we investigate the most commonly used approaches, their application, and their other uses beyond evaluating models. The survey concludes by highlighting some promising directions for future research.

1 Introduction

The performance of natural language generation (NLG) technology (Reiter, 2024) has improved dramatically in recent years (Clark et al., 2021). As the errors made by NLG systems are becoming subtler, the detection of errors in NLG models is likewise becoming more complex. Consequently, recent years have seen a growing focus on the *evaluation* of NLG models (Sai et al., 2022; Celikyilmaz et al., 2020; Novikova et al., 2017; Li et al., 2024).

The evaluation of NLG models can be categorized into human and automatic evaluation. Although human evaluation is widely regarded as more convincing than computational metrics (van der Lee et al., 2019; Graham et al., 2017), human evaluation tends to be costly. Therefore, automatic evaluation is often used in practice. Automatic evaluation can be divided into two types: reference-based and reference-free. Reference-based metrics evaluate the quality of a text by measuring the correspondence between system outputs and human-written texts, termed references,

which are considered gold standards for evaluation. Reference-based metrics require the preparation of references, and humans must create and verify the reference texts. In addition, many NLG tasks require multiple reference texts because the same information may be expressed in many different ways, no one of which is necessarily better than the others. Moreover, the performance of reference-based methods greatly depends on the quality and quantity of the references (Freitag et al., 2020). If there are few references, or their quality is poor, they will not provide proper evaluation.

To address the limitations of reference-based metrics, reference-free metrics, which evaluate NLG systems without references, have been proposed.¹ Reference-free evaluation has the potential to greatly increase the scalability of NLG evaluation. Moreover, the performance of reference-free evaluation, in terms of the strengths of its correlation with human evaluation ratings, has recently seen improvements (Rei et al., 2021; Islam and Magnani, 2021).

While many reference-free approaches have been proposed, most NLG evaluation studies have so far been reference-based. Unsurprisingly therefore, surveys of evaluation metrics for individual NLG tasks (Chauhan and Daniel, 2022; Ermakova et al., 2019) and comprehensive surveys of NLG evaluations (Celikyilmaz et al., 2020; Sai et al., 2022) have paid little attention to reference-free approaches. The present survey intends to fill this gap, paying attention to the full breadth of NLG tasks (Table 1). To clarify the value and limitations of reference-free evaluation metrics and to encourage their future study, we offer a survey of the main approaches to reference-free evaluation (e.g., Chimoto and Bassett, 2022) and their analysis (e.g., Mohiuddin et al., 2021; Durmus et al., 2022).

¹Reference-free evaluation is also called quality estimation (QE) (Callison-Burch et al., 2012; Scarton et al., 2016).

2 Terminology

In this paper, we use the term **reference** (r) for a piece of text that can be considered to be correct, for instance because it was written or validated by humans. References are used for assessing the quality of the outputs of the NLG model. Outputs are thus viewed as **hypotheses** (h). All information used for the evaluation metrics other than hypotheses and references is **context** (c). Often, the context is the NLG model’s input, but not always. For example, in a dialogue generation task, some evaluation metrics make use of future utterances, which the NLG model does not have access to (Li et al., 2021; Mehri and Eskenazi, 2020a).

Reference-free metrics are typically categorized into two types: absolute evaluation and ranking evaluation. Absolute evaluation takes a context (c) and a hypothesis (h), and returns a quality score. Ranking evaluation takes a context (c) and a set of hypotheses, and ranks the hypotheses; in particular, the method of ranking two hypotheses (h_1 and h_2) is called pairwise evaluation.

Table 1 presents contexts and hypotheses for each NLG task. This table is reconstructed from Table 2 in Sai et al. (2022), adding the story generation task and revising some of the task settings. Unless otherwise noted, these formats of contexts and hypotheses for each task are assumed in this paper.

3 Research questions

Generally, reference-based approaches use similarity to a reference to estimate the quality of the hypothesis, given the context; we will say that such similarity metrics function as a *proxy* for the quality of the hypothesis (given the context). In the absence of references, text quality assessment methods differ along a number of important dimensions, which will be used to organize this survey.

The first question is (i) *How do these approaches evaluate the quality of the hypothesis, and what do they use as a proxy for quality?* (Section 4) While many methods consider both the context and the hypothesis, some focus exclusively on evaluating the hypothesis alone. Often used supplementary in reference-free evaluation scenarios, these methods offer an approach to assessing textual qualities such as fluency. This raises the question (ii) *How can hypothesis-only evaluation methods evaluate textual qualities such as fluency?* (Section 5) Finally, reference-free metrics can have other uses

beyond comparing NLG models’ performance. For example, they can be used in a re-ranking function to select the most optimal hypotheses from an NLG model and in a reward function for reinforcement learning.

They have also been used to filter out low-quality training data of NLG models (Bane et al., 2022). We therefore ask, (iii) *How can reference-free metrics be used, other than for evaluating the quality of the hypothesis in an NLG task?* (Section 6).

4 What is the proxy for quality?

A reference-based method evaluates text quality by comparing the text (i.e., the hypothesis) with the reference. For reference-free approaches, we discuss how quality is evaluated and what is used as an evaluation proxy. Some reference-free metrics combine several of the approaches presented below. Methods that focus solely on the hypothesis, ignoring the context, are discussed in Section 5.

4.1 Learning from human judgments

One approach to constructing reference-free metrics is to use a regression model to predict human judgments given a context and a hypothesis (Heilmann et al., 2014; Xenouleas et al., 2019). This idea is straightforward because if an evaluation metric correlates strongly with human judgments, it must be reliable. For example, COMET-QE (Rei et al., 2021) uses a regression model trained on human scores of MT using features of the context and the hypotheses that are acquired using a cross-lingual pre-trained language model.

Scope and merits. This method can be applied to all NLG tasks (Table 1), as long as high-quality human judgments are available in large quantities. In particular, in the context of MT evaluation, human evaluation data is consistently compiled for the annual shared task of evaluation metrics (Freitag et al., 2021, 2022, 2023), which has led to the frequent application of this method in MT (Rei et al., 2021). With a large quantity of high-quality human judgments, it should be possible to construct high-performance evaluation metrics in this way.

Limitation. Collecting large amounts of high-quality human judgments is costly.

4.2 Learning from pseudo-judgments

Instead of human judgments, pseudo-judgments are likewise used as a proxy. The most prev-

Task	Context	Hypothesis
Machine Translation (MT)	Source language text	Translation
Summarization (SUM)	Document(s)	Summary
Question Answering (QA)	Question + Background knowledge (e.g., Knowledge base, Image)	Answer
Question Generation (QG)	Background knowledge (e.g., Knowledge base, Image)	Question
Dialogue (DG)	Conversation history	Response
Story Generation (SG)	Premises or Summary	Story
Image Captioning (IC)	Image	Caption
Data-to-Text (D2T)	Structured data (e.g., Table, Graph)	Text

Table 1: Context and hypothesis of typical NLG tasks, modified from Table 2 in [Sai et al. \(2022\)](#).

lent method for generating pseudo-ratings involves utilizing scores from reference-based evaluation metrics ([Chollampatt and Ng, 2018](#); [Zouhar et al., 2023](#)). In this case, generating pseudo-judgments requires reference and hypothesis data, which are often NLG model training data and the corresponding output from an NLG model, respectively.

To generate pseudo-judgments, simple approaches are also sometimes used. For example, the reference is given a “good” judgment, while the output of the NLG model or a noised version of the reference (such as with word drops) is assigned a “bad” judgment ([Bao et al., 2022](#); [Guan and Huang, 2020](#); [Wu et al., 2020](#); [Lee et al., 2021b](#); [Moosa et al., 2024](#)). This approach could be better suited for building pairwise ranking evaluation metrics because a simple “good or bad” decision would be sufficient for constructing such metrics, and it makes it easier to generate training data for pairwise comparisons ([Moosa et al., 2024](#)).

Scope and merits. This approach is used for data augmentation to supplement limited human judgments. Moreover, this approach is used to effectively transforms reference-based evaluation methods into reference-free evaluation metrics, allowing for the benefits of reference-free methods, as shown in Section 6.

Limitation. It is important to note that discrepancies between pseudo-judgments and actual evaluations are likely.

4.3 Correspondence between context and hypothesis

“Correspondence” is the extent to which the hypothesis aligns with the context. In various NLG tasks, the hypothesis should faithfully reflect the context and accurately convey the necessary details. Therefore, the correspondence between the context and hypothesis is often used as a proxy for evaluating these tasks. We discuss the general scope, mer-

its, and limitations of this approach, followed by a detailed explanation of the method for comparing context and hypotheses.

Scope and merits. The correspondence-based approach is often used in machine translation, summarization tasks, image captioning, and data-to-text tasks ([Gatt and Krahmer, 2018](#); [Reiter, 2024](#)). For example, in machine translation, the output must express the same information as the input; in summarization, the emphasis is on whether the generated summary encapsulates the most important information from the original text; similarly, in image captioning, the emphasis is on accurately describing the main content of the image. Information not present in the context should not appear in the hypothesis. The correspondence-based approach is used to test these aspects. The correspondence-based approach is applied to test the veracity of the hypothesis ([van Deemter, 2024](#)), as it is commonly employed to detect hallucinations in generated content. In particular, approaches using Question Answering (QA) and Natural Language Inference (NLI) are often used for this purpose ([Fabbri et al., 2022](#)), as will be explained below.

Limitations. Although these methods can successfully apply to the veracity of generated texts, they are difficult to apply to such quality issues as verbosity, duplication of information, lack of coherence, lack of fluency, and impoliteness. Other limitations apply to other NLG tasks. For example, in story generation, new information, which is not present in the context, often has to be generated, and the correspondence-based approach is not suitable for evaluating the quality of such new information. The correspondence-based approach can only test the presence of information that should (or should not) be included in the hypothesis based on the context. Therefore, it is often used in combination with other approaches. For example, in machine translation, it is used in conjunction with

the perplexity of hypotheses as given by language models (Zhao et al., 2020).

4.3.1 Encoding as embedding

One way to perform correspondence-based evaluation involves encoding both the context and the hypothesis as embeddings, and then assessing the relationship using measures such as cosine similarity or word mover’s distance.

Scope and merits. The method is often used in machine translation and image captioning. XMoverScore (Zhao et al., 2020), used in machine translation, and CLIPScore (Hessel et al., 2021), used in image captioning, are examples of evaluation metrics that use this approach. XMoverScore utilizes multilingual BERT (MBERT) (Devlin et al., 2019), which supports multilingual text embeddings, while CLIPScore utilizes CLIP (Radford et al., 2021), which can encode both images and text as embeddings.

Limitations. An embedding model that consistently places corresponding context and hypothesis close in its vector space is important for this approach. However, not all embedding models ensure this. For example, it has been suggested that the monolingual subspaces of MBERT, which is used as an embedding for machine translation evaluation metrics, do not align well with each other, and require re-mapping (Zhao et al., 2020; Belouadi and Eger, 2022).

4.3.2 Reverse transformation.

Another approach is to follow up the transformation performed by NLG by its reverse transformation, then compare the original context with the result of that reverse transformation, then evaluate whether the original context has been correctly restored.

Scope and merits. This approach to correspondence-based evaluation is often used in machine translation, where back-translation models are readily available. For example, in English-to-German translation (forward translation), a German-to-English translation model is prepared, and then the input English text of the model is compared with the back-translated English text, using reference-based metrics (Moon et al., 2020; Zhuo et al., 2023). When using an NLG model for the reverse transformation, it is similar to methods that will be discussed in Section 4.4.2. In other tasks, such as generating text from semantic representations, Manning and

Schneider (2021) a similar approach has been applied by using a parser that converts text to semantic representations to evaluate whether the generation process can be accurately reversed.

Limitations. The main limitation of this approach is that it depends on the performance of the reverse transformation (Moon et al., 2020; Zhuo et al., 2023).

4.3.3 Using question answering (QA).

QA is used to evaluate whether a text contains the necessary information. The idea behind this approach is to check whether the answers obtained by a QA system, when referencing the context and the hypothesis separately for a given question, are the same. If the answers are same, it is assumed that the information regarding the question is contained in both the context and the hypothesis. The critical point in this approach is how to extract suitable questions.

Scope and merits. This approach is often applied for the evaluation of summarization. Eyal et al. (2019) proposed a (reference-based) method for identifying important entities from references and generating fill-in-the-blank questions. Then, Scialom et al. (2019) extended the method to generate fill-in-the-blank questions from the context, and Wang et al. (2020) generated questions from the hypothesis. Scialom et al. (2021) proposed a method for generating questions from both context and hypothesis. The idea of using QA for testing other NLP systems was used by Lee et al. (2021a) in testing a caption generation system and by Rebuffel et al. (2021) in testing a table-to-text system.

Limitations. The success of this approach depends on two elements: (i) creating good questions (Gabriel et al., 2021) and (ii) having a high-performing QA system.

4.3.4 Using natural language inference (NLI).

NLI is the task of predicting whether one sentence or text implies or contradicts another (or neither, i.e., “neutral”). NLI-based approach are often used for summarization (Laban et al., 2022) and dialogue evaluation (Dziri et al., 2019; Pang et al., 2020). For example, in dialogue, it should be noted that each utterance is consistent with previous ones,², and NLI is often used to detect incon-

²Recall that the *context* of an utterance in dialogue is considered to consist of the utterances preceding it in the dialogue. (See Table 1)

sistencies.

Scope and merits. NLI-based approaches are often used for summarization (Laban et al., 2022) and dialogue evaluation (Dziri et al., 2019; Pang et al., 2020). For example, in dialogue, it should be noted whether each utterance is consistent with previous ones,³, and NLI is often used to detect inconsistencies.

Limitations. A simple way to perform this approach is to use models trained on existing NLI datasets, such as SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). However, these existing datasets are known to have biases (Gururangan et al., 2018). In addition, existing NLI datasets may not meet the requirements for NLG evaluation due to factors such as domain mismatch. Indeed, Falke et al. (2019) confirmed that models trained on existing NLI datasets cannot be used immediately for summarization evaluation. Dziri et al. (2019) built NLI models by creating a pseudo-NLI dataset for dialogue with normal conversational data as entailment, unrelated dialogue and dull responses as neutral, and ungrammatical sentences and contradiction data of MultNLI as a contradiction.

4.3.5 Comparing key expressions.

Lastly, the correspondence-based approach can be performed by extracting key expressions from the context and hypothesis and comparing them. The critical point in this relatively simple approach is how to extract keywords. For example, KoBE (Gekhman et al., 2020), a metric for evaluating machine translation, uses an entity linking system with a multilingual knowledge base to extract entities (keywords) from both the context and hypothesis, and then compares these entities to check correspondences. In image captioning, Madhyastha et al. (2019) proposed a metric that detects an object in an image and compares the labels of detected objects, and the caption using the Word Mover Distance.

Scope and merits. This method is good for testing whether or not a certain thing is mentioned. It is a simple method that is useful for checking for obvious omissions.

³Recall that the *context* of an utterance in dialogue is considered to consist of the utterances preceding it in the dialogue. (See Table 1)

Limitations. This approach is intrinsically very limited because keyword extraction does not address the full meaning of the hypothesis or the context. For example, even if the text mentions all the correct objects, adding a negation will change the meaning of the sentence dramatically, potentially turning a good hypothesis into a bad one or conversely.

4.4 Peer evaluation

Frequently, one NLG model is used as a proxy for evaluating another. With some irony, this had been called “peer” evaluation.

4.4.1 Probability given other NLG models

In order to address the correspondence between context and hypothesis, high-performing NLG models that differ from the model to be evaluated are often used (Fu et al., 2023; Thompson and Post, 2020). For example, Prism-src (Thompson and Post, 2020) uses a multilingual machine translation model to calculate the generation probability of the hypothesis for the context and take that probability as the evaluation score. In addition, Mehri and Eskenazi (2020a); Li et al. (2021) propose the use of a dialogue model to evaluate other dialogue models.

Scope and merits. This method can be applied to all NLG tasks (Table 1) in principle, as long as a high-performance NLG model is available.

Limitations. This approach suffers from the issue that a high-performance NLG model is necessary to evaluate the performance of an NLG model (Deutsch et al., 2022). Meanwhile, Agrawal et al. (2021) suggest that an evaluation translation model can correctly rank other translation models that outperform it on average. Further investigation is needed to weigh the strengths and weaknesses of this approach.

4.4.2 Similarity with pseudo-reference

The idea of this approach is to use pseudo-references (Belouadi and Eger, 2022; Chen et al., 2021; Gao et al., 2020). Here, Similarity is usually calculated by a reference-based metric. This approach may be regarded as one of the methods outlined in Section 4.3, as it addresses the challenges of context and hypothesis formulation by employing pseudo-references.

Scope and merits. Pseudo-references are generated by NLG models other than the model being evaluated or, in some cases, by heuristic algorithms.

Therefore, the method can be applied to all NLG tasks (Table 1), as long as a high-performance NLG model is available. For example, Belouadi and Eger (2022) propose an evaluation metric for machine translation, part of which creates pseudo-references using another machine translation model and compares the pseudo-references to the output of the test model, using Word Mover Distance.

Limitations. Analogous to the reference-based approach, the quality of the pseudo-references significantly influences the evaluation performance. A notable challenge for this approach is the necessity of having access to a high-performance NLG model for the effective evaluation of other NLG models.

4.4.3 LLM-as-a-judge

Recent advancements have seen the increasing utilization of LLMs in the evaluation of NLG tasks, a method commonly referred to as “LLM-as-a-judge” (Zheng et al., 2023; Chiang and Lee, 2023; Kocmi and Federmann, 2023). The LLM-as-a-judge presents evaluation results as text output, such as generating evaluation scores like “this hypothesis is 3 score”. The expression in the formula is as follows.

Scope and merits. The advantage of LLMs lies in its ability to adapt to various tasks and evaluation criteria through prompting. Additionally, LLMs can generate explanation such as the reasoning behind scores. Furthermore, the emergence of multi-modal LLMs has shown promising results in tasks such as image caption (Lee et al., 2024).

Limitations. While it is a promising approach, there are several challenges (Zheng et al., 2023; Chen et al., 2024; Ohi et al., 2024). For example, it has been reported that slight modifications in the prompt can alter evaluations, and there are biases like giving higher ratings to longer hypotheses (Zheng et al., 2023). Additionally, most LLMs used in this approach are proprietary commercial models (e.g., GPT-4), only accessible via an API, and subject to frequent updates, which poses challenges to reproducibility.

5 How can textual quality be evaluated?

Some reference-free metrics are designed to evaluate hypotheses only, without paying attention to context. Such metrics are used to evaluate textual quality and they are often combined with other

reference-free metrics, such as those described in Section 4.3. This section focuses on often discussed textual qualities, such as fluency, and on typical approaches to using these qualities in NLG evaluation.

5.1 Supervised modeling of textual quality on annotated data

Research on tasks such as essay scoring, which involve qualities such as fluency and coherence, is actively conducted. Models trained for these tasks are also used as evaluation metrics in NLG. Here, we briefly discuss several datasets in which scores for fluency and coherency have been manually assigned to texts. Datasets of this kind can be used for training models of textual quality, offering significant benefits. However, the construction of such datasets involves considerable expense.

Fluency. The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) is a dataset comprising English sentences, each annotated as either grammatical or ungrammatical. CoLA is often used to train classification models for fluency assessment (Zhu and Bhat, 2020; Krishna et al., 2020). Italian CoLA (Trotta et al., 2021) and Russian CoLA (Mikhailov et al., 2022) have also been created. SummEval (Fabbri et al., 2021) is a benchmark dataset for summarization evaluation, featuring fluency score annotations applied to the outputs of various summarization models. Additionally, TMU-GFM-Dataset (Yoshimura et al., 2020) is a dataset specifically designed for Grammatical Error Correction (GEC) metrics. This dataset includes annotations for both fluency and grammaticality in the outputs of GEC systems.

Coherence. The Grammarly Corpus of Discourse Coherence (GCDC) (Lai and Tetreault, 2018) is a dataset featuring English texts that have been manually annotated to reflect three levels of discourse coherence: low, medium, and high. Classification model trained on GCDC is used for evaluating coherence (Vásquez-Rodríguez et al., 2023). Following the GCDC, datasets annotated for discourse coherence in Danish (Flansmose Mikkelsen et al., 2022) and Chinese (Wu et al., 2023) have subsequently been created. SummEval also provides coherence score as well as fluency.

5.2 Language models

LMs are often used to evaluate the fluency, coherence, and readability of hypotheses, without rely-

ing on annotated data, without relying on annotated data (Mehri and Eskenazi, 2020b; Wu et al., 2020). Texts with high probability given an LM are considered high-quality texts. In practice, several methods can be used to calculate the score, such as perplexity, negative log-likelihood, and SLOR (Kann et al., 2018). For example, the Scribendi Score (Islam and Magnani, 2021), a reference-free evaluation metric for GEC based on GPT-2’s perplexity scores, has been proposed. This metric is reported to correlate better with human ratings than traditional reference-based metrics such as M2 (Dahlmeier and Ng, 2012) and GLEU (Napoles et al., 2015, 2016). This approach is often combined with others, such as those described in Section 4.3.

5.3 Approaches focused on individual textual quality aspects

This subsection discusses several approaches that address specific aspects of textual quality.

5.3.1 Coherence

Coherence evaluation methods are often trained on synthetic tasks. For example, a synthetic task is a task that creates incoherent text (negatives) by inserting extra sentences or shuffling the order of sentences, and then classifies negatives and positives (Moon et al., 2019; Shen et al., 2021).

These tasks are similar to Next Sentence Prediction and Sentence Order Prediction used in the pre-training phase of masked language models such as BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2020). Therefore, masked language models are also used to evaluate text coherence (Zhu and Bhat, 2020).

However, it has been reported that even if the performance is high on synthetic tasks, performance on downstream NLG tasks can be low (Mohiuddin et al., 2021). Steen and Markert (2022) also investigate the performance of reference-free coherence evaluation metrics in summarization tasks and found that the correlation between human judgment and metrics scores was low.

5.3.2 (Lack of) redundancy

Lack of redundancy is an important quality criterion for tasks generating paragraph- and document-level text, such as summarization and story generation. To measure redundancy, heuristic methods such as n-gram repetition counts and superficial expression repetition counts are commonly used (Zhu

and Bhat, 2020; Xiao and Carenini, 2020). However, these methods fail to capture contextual redundancies, where different surface forms have semantically similar content.

5.3.3 Readability

Readability assessment is the task of estimating how easy or difficult a text is to read, which has long been addressed, not only in the language processing but also in language education. Fluency, redundancy, and coherence, among other factors, contribute to readability. Another important aspect of readability relates to its readers, because it is thought that expressions that are unnecessarily difficult for readers should be avoided.

Readability measures, including and Flesch Reading Ease and Flesch-Kincaid Grade Level (Flesch, 1948; Kincaid et al., 1975), have long been used. These measures are linear combinations of the number of words per sentence, the number of syllables per word, and so on, using carefully adjusted weights. Flesch-Kincaid Grade Level is often used in evaluation of simplification tasks (Alva-Manchego et al., 2019). However, it is reported that these methods can yield inaccurate or misleading results (Tanprasert and Kauchak, 2021), so caution is advised in their utilization.

5.3.4 Diversity

Diversity is often considered to be desirable in tasks such as dialogue, story generation, and paraphrasing. Most diversity evaluation metrics are surface based, based on n-grams; often used ones are Distinct-N (Li et al., 2016), Self-BLEU (Sun and Zhou, 2012; Serban et al., 2017)⁴, and Pairwise-BLEU (Shen et al., 2019). In practice, increasing diversity may sacrifice the clarity of the generated text, so it should always be used in conjunction with other evaluation criteria (Ippolito et al., 2019; Kulikov et al., 2019).

6 Other uses of reference-free metrics

We discuss other uses of reference-free metrics that go beyond their use in evaluating models.

6.1 Rewards for reinforcement learning

The reference-free approach is often used as a reward function for reinforcement learning (Scialom

⁴There are two methods called Self-BLEU; one computes the BLEU score between inputs and outputs (Sun and Zhou, 2012) and the other computes the BLEU score between generated sentence sets (Serban et al., 2017).

et al., 2019; Cho et al., 2022). A reference-based metric is also sometimes used as a reward function, but it cannot be used online and is typically used to optimize a model with the objective function (evaluation metrics) for each task (Bahdanau et al., 2017).

Reinforcement learning from human feedback (RLHF), which optimizes NLG models based on human feedback on their output to reflect human preference, has gained much attention (Christiano et al., 2017; Stiennon et al., 2020). In the RLHF framework, a model that outputs a scalar reward to quantify human preferences is used as the reward function, not human feedback directly. This reward model is basically the same as that of the reference-free metrics introduced in Section 4.1. Thus, improvement to reference-free evaluation metrics has become increasingly important for both evaluating and learning NLG models.

6.2 Data selection for training NLG models

Huge, high-quality data can help train better NLG models (Kaplan et al., 2020). In recent years, approaches to crawling data on the Web (Bañón et al., 2020) and various data augmentation techniques (Feng et al., 2021) have been developed to increase the training data for NLG models. These techniques can increase the training data, but these augmented data generally contain noise. Therefore, it is important to use cleaning technology. The data-cleaning task is essentially the same as the reference-free evaluation metric. For example, dual conditional cross-entropy filtering (Junczys-Dowmunt, 2018) is essentially the same idea as the method introduced in Section 4.4.1.

In active learning, a reference-free evaluation metric is also used to select data to be requested for humans to annotate (Chimoto and Bassett, 2022; Zeng et al., 2019). By requesting humans to annotate the output of an NLG model with low scores on the reference-free metrics can produce data to build better NLG models that are preferentially selected and labeled.

Studies of data selection can help identify some weaknesses of reference-free metrics, as they apply these metrics to different data distributions than those typically used in the evaluation of NLG evaluation metrics. In fact, Bane et al. (2022) compared some data filtering methods to find that COMET-QE is strong in detecting word order but not in detecting mismatches of numbers between context and target text. Complementary research between

data selection and reference-free metrics is becoming increasingly important.

6.3 Reranking and filtering of model outputs

Reranking or filtering relates to the task of selecting good outputs or removing bad outputs from a set of hypothesis. Reference-free evaluation metrics can be used for these tasks (Chollampatt and Ng, 2018). In comparison to standard model evaluation, reranking and filtering require often runtime evaluation, which need to be done quickly. Therefore, efficient reference-free evaluation metrics (Grünwald et al., 2022; Zhang et al., 2022) are valuable for these functions.

7 Discussion

7.1 Reference-free vs. reference-based

In this section, we summarize the main characteristics of reference-free and reference-based metrics.

Faced with the choice between reference-based and reference-free methods, an important consideration is related to the *cost* of the method. Reference-based approaches require references, which are often expensive to collect. In particular, it is not easy to collect enough references for open-ended tasks, such as dialogue and story generation. Reference-free metrics do not require references. However, as pointed out in Section 4, many reference-free methods require training data, such as human judgments, or parallel data. Consequently, a reference-free approach is by no means cost-free.

The second consideration is *performance*. Some researchers have reported that reference-free metrics produce lower evaluation performance than reference-based ones (Banchs and Li, 2011; Fonseca et al., 2019), but in recent years, some studies have reported the reverse. For example, Kasai et al. (2022) found that reference-free metrics perform better than reference-based ones when there are few references. If the reference quality is poor, the evaluation performance of the reference-based method suffer (see Table 13 of Freitag et al. (2021)). Reference-free metrics are thus not always inferior to reference-based metrics.

7.2 LLMs and reference-free evaluation metrics

LLMs have altered the NLG evaluation landscape in broadly two ways.

On the one hand, LLMs have improved the performance of NLG models. As a result, texts gener-

ated by these NLG models are often indistinguishable from those written by humans (Clark et al., 2021). These improvements present a challenge for the evaluation of such texts because, in some cases, the errors and infelicities that separate competing models are becoming increasingly subtle.

On the other hand, with the advent of LLMs, reference-free metrics have also advanced. For example, in addition to the methods presented in Section 4.4.3, LLMs are used as regression models in Section 4.1 (Guerreiro et al., 2023). It would also be possible to use LLMs as NLG models in the methods presented in Section 4.4.2 and 4.4.1 in the creation of pseudo-ratings presented in Section 4.2.⁵

7.3 Future research: Combining different approaches to evaluation

In future, we expect that different approaches to evaluation, which have complementary strengths, will be combined to achieve superior results. For example, it has been suggested that the types of errors that humans can easily detect differ from those that automatic evaluation metrics can easily identify (Chen et al., 2024). By combining human and automatic evaluation, it may therefore be possible to reduce the workload of human evaluators (Zhang et al., 2021), or to achieve results with superior validity and reliability. For example, human evaluation could focus on cases in which discrepancies occur across evaluation metrics. We believe that exploring effective and efficient ways to combine human evaluation and automatic evaluation metrics is an important direction for future research.

8 Conclusion

Evaluation is an essential step for the development of NLG models of any kind(Gatt and Krahmer, 2018; van der Lee et al., 2019; Reiter, 2024). In this survey, we have discussed existing computational, *reference-free* evaluation metrics for a broad range of NLG tasks, including not only data-to-text NLG but also such text-to-text tasks as machine translation and question answering, for example, which are not always included in discussions of NLG and NLG evaluation. These reference-free metrics have recently attracted more attention and shown improved performance.

We have shown that although a wide range of reference-free metrics have been proposed, all of these can be seen as variations on a few basic themes, such as learning from human judgments; exploiting the correspondence between the context and the hypothesis (as these notions are defined in section 2); and using peer evaluation. We have also outlined the assumptions and limitations underlying each method.

We hope that this survey will lead to further research on reference-free metrics, and that this will lead to further insights in the question of when these metrics are most useful, and how they are best combined with other approaches to evaluation.

⁵See the survey by Li et al. (2024) for evaluation metrics using LLMs.

References

- Sweta Agrawal, George Foster, Markus Freitag, and Colin Cherry. 2021. [Assessing Reference-Free Peer Evaluation for Machine Translation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1171, Online. Association for Computational Linguistics.
- Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. [EASSE: Easier Automatic Sentence Simplification Evaluation](#). 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): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.
- Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. [An Actor-Critic Algorithm for Sequence Prediction](#). In *International Conference on Learning Representations*.
- Rafael E. Banchs and Haizhou Li. 2011. [AM-FM: A Semantic Framework for Translation Quality Assessment](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 153–158, Portland, Oregon, USA. Association for Computational Linguistics.
- Fred Bane, Celia Soler Uguet, Wiktor Stribizew, and Anna Zaretskaya. 2022. [A Comparison of Data Filtering Methods for Neural Machine Translation](#). In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 313–325, Orlando, USA. Association for Machine Translation in the Americas.
- Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. [ParaCrawl: Web-Scale Acquisition of Parallel Corpora](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.
- Forrest Bao, Ge Luo, Hebi Li, Minghui Qiu, Yinfei Yang, Youbiao He, and Cen Chen. 2022. [SueNes: A Weakly Supervised Approach to Evaluating Single-Document Summarization via Negative Sampling](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2450–2458, Seattle, United States. Association for Computational Linguistics.
- Jonas Belouadi and Steffen Eger. 2022. [USCORE: An Effective Approach to Fully Unsupervised Evaluation Metrics for Machine Translation](#). *arXiv preprint*.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. [Findings of the 2012 Workshop on Statistical Machine Translation](#). In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. [Evaluation of Text Generation: A Survey](#). *arXiv preprint*.
- Shweta Chauhan and Philemon Daniel. 2022. [A Comprehensive Survey on Various Fully Automatic Machine Translation Evaluation Metrics](#). *Neural Processing Letters*, pages 1–55.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. [Humans or LLMs as the Judge? A Study on Judgement Bias](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.
- Wang Chen, Piji Li, and Irwin King. 2021. [A Training-free and Reference-free Summarization Evaluation Metric via Centrality-weighted Relevance and Self-referenced Redundancy](#). 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 404–414, Online. Association for Computational Linguistics.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can Large Language Models Be an Alternative to Human Evaluations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Everlyn Chimoto and Bruce Bassett. 2022. [COMET-QE and Active Learning for Low-Resource Machine Translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4735–4740, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. [Fine-grained Image Captioning with CLIP Reward](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 517–527, Seattle,

- United States. Association for Computational Linguistics.
- Shamil Chollampatt and Hwee Tou Ng. 2018. [Neural Quality Estimation of Grammatical Error Correction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2528–2539, Brussels, Belgium. Association for Computational Linguistics.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. [Deep Reinforcement Learning from Human Preferences](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- 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.
- Daniel Dahlmeier and Hwee Tou Ng. 2012. [Better Evaluation for Grammatical Error Correction](#). In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.
- Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. [On the Limitations of Reference-Free Evaluations of Generated Text](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. 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.
- Esin Durmus, Faisal Ladakh, and Tatsunori Hashimoto. 2022. [Spurious Correlations in Reference-Free Evaluation of Text Generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1443–1454, Dublin, Ireland. Association for Computational Linguistics.
- Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. [Evaluating Coherence in Dialogue Systems using Entailment](#). 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 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics.
- Liana Ermakova, Jean Valère Cossu, and Josiane Mothe. 2019. [A survey on evaluation of summarization methods](#). *Information Processing & Management*, 56(5):1794–1814.
- Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. [Question Answering as an Automatic Evaluation Metric for News Article Summarization](#). 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 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. [QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [SummEval: Re-evaluating Summarization Evaluation](#). *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. [Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.
- Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. [A Survey of Data Augmentation Approaches for NLP](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.
- Linea Flansmose Mikkelsen, Oliver Kinch, Anders Jess Pedersen, and Ophélie Lacroix. 2022. [DDisCo: A discourse coherence dataset for Danish](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2440–2445, Marseille, France. European Language Resources Association.
- Rudolph Flesch. 1948. [A new readability yardstick](#). *Journal of Applied Psychology*, 32(3):221.
- Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. [Findings of the WMT 2019 Shared Tasks on Quality Estimation](#). In *Proceedings of the Fourth Conference*

- on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.
- Markus Freitag, David Grangier, and Isaac Caswell. 2020. **BLEU might be Guilty but References are not Innocent**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.
- Markus Freitag, Nitika Mathur, Chi-ku Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. **Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent**. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-ku Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. **Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust**. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-ku Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. **Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain**. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. **GPTScore: Evaluate as You Desire**. *Preprint*, arXiv:2302.04166.
- Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. **GO FIGURE: A Meta Evaluation of Factuality in Summarization**. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.
- Yang Gao, Wei Zhao, and Steffen Eger. 2020. **SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.
- 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.
- Zorik Gekhman, Roee Aharoni, Genady Beryozkin, Markus Freitag, and Wolfgang Macherey. 2020. **KoBE: Knowledge-Based Machine Translation Evaluation**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3200–3207, Online. Association for Computational Linguistics.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. **Can machine translation systems be evaluated by the crowd alone**. *Natural Language Engineering*, 23(1):3–30.
- Jens Grünwald, Christoph Leiter, and Steffen Eger. 2022. **Can we do that simpler? Simple, Efficient, High-Quality Evaluation Metrics for NLG**. *arXiv preprint*.
- Jian Guan and Minlie Huang. 2020. **UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166, Online. Association for Computational Linguistics.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. **xcomet: Transparent machine translation evaluation through fine-grained error detection**. *Preprint*, arXiv:2310.10482.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. **Annotation Artifacts in Natural Language Inference Data**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. **Predicting Grammaticality on an Ordinal Scale**. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180, Baltimore, Maryland. Association for Computational Linguistics.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. **CLIPScore: A Reference-free Evaluation Metric for Image Captioning**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. **Comparison of Diverse Decoding Methods from Conditional Language Models**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.
- Md Asadul Islam and Enrico Magnani. 2021. **Is this the end of the gold standard? A straightforward**

- reference-less grammatical error correction metric. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3009–3015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Marcin Junczys-Dowmunt. 2018. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.
- Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-Level Fluency Evaluation: References Help, But Can Be Spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, Brussels, Belgium. Association for Computational Linguistics.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint*.
- Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander Fabbri, Yejin Choi, and Noah A. Smith. 2022. Bidimensional Leaderboards: Generate and Evaluate Language Hand in Hand. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3540–3557, Seattle, United States. Association for Computational Linguistics.
- J Peter Kincaid, Robert P Fishburne Jr, Richard 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.
- Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating Unsupervised Style Transfer as Paraphrase Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.
- Ilya Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of Search and Evaluation Strategies in Neural Dialogue Modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Alice Lai and Joel Tetreault. 2018. Discourse Coherence in the Wild: A Dataset, Evaluation and Methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.
- Hwanhee Lee, Thomas Scialom, Seunghyun Yoon, Franck Dernoncourt, and Kyomin Jung. 2021a. QACE: Asking Questions to Evaluate an Image Caption. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4631–4638, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021b. UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 220–226, Online. Association for Computational Linguistics.
- Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3732–3746, Bangkok, Thailand. Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances. 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 128–138, Online. Association for Computational Linguistics.

- Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging Large Language Models for NLG Evaluation: A Survey. *arXiv preprint arXiv:2401.07103*.
- Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2019. **VIFIDEL: Evaluating the Visual Fidelity of Image Descriptions**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550, Florence, Italy. Association for Computational Linguistics.
- Emma Manning and Nathan Schneider. 2021. **Referenceless Parsing-Based Evaluation of AMR-to-English Generation**. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 114–122, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Shikib Mehri and Maxine Eskenazi. 2020a. **Unsupervised Evaluation of Interactive Dialog with Di-aloGPT**. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.
- Shikib Mehri and Maxine Eskenazi. 2020b. **USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.
- Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. **RuCoLA: Russian Corpus of Linguistic Acceptability**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tasnim Mohiuddin, Prathyusha Jwalapuram, Xiang Lin, and Shafiq Joty. 2021. **Rethinking Coherence Modeling: Synthetic vs. Downstream Tasks**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3528–3539, Online. Association for Computational Linguistics.
- Han Cheol Moon, Tasnim Mohiuddin, Shafiq Joty, and Chi Xu. 2019. **A Unified Neural Coherence Model**. 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 2262–2272, Hong Kong, China. Association for Computational Linguistics.
- Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020. **Revisiting Round-trip Translation for Quality Estimation**. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.
- Ibraheem Muhammad Moosa, Rui Zhang, and Wenpeng Yin. 2024. **MT-Ranker: Reference-free machine translation evaluation by inter-system ranking**. In *The Twelfth International Conference on Learning Representations*.
- Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. **Ground Truth for Grammatical Error Correction Metrics**. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. Association for Computational Linguistics.
- Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2016. **GLEU without tuning**. *arXiv preprint arXiv:1605.02592*.
- 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.
- Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki. 2024. **Likelihood-based Mitigation of Evaluation Bias in Large Language Models**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3237–3245, Bangkok, Thailand. Association for Computational Linguistics.
- Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. **Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. **Learning Transferable Visual Models From Natural Language Supervision**. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. **Data-QuestEval: A Referenceless Metric for Data-to-Text Semantic Evaluation**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8029–8036, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya

- Glushkova, André F. T. Martins, and Alon Lavie. 2021. [Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.
- Ehud Reiter. 2024. *Natural Language Generation*. Springer.
- 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).
- Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2016. [Quality Estimation for Language Output Applications](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 14–17, Osaka, Japan. The COLING 2016 Organizing Committee.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. [QuestEval: Summarization Asks for Fact-based Evaluation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. [Answers Unite! Unsupervised Metrics for Reinforced Summarization Models](#). 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 3246–3256, Hong Kong, China. Association for Computational Linguistics.
- Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. [A hierarchical latent variable encoder-decoder model for generating dialogues](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).
- Aili Shen, Meladel Mistica, Bahar Salehi, Hang Li, Timothy Baldwin, and Jianzhong Qi. 2021. [Evaluating Document Coherence Modeling](#). *Transactions of the Association for Computational Linguistics*, 9:621–640.
- Tianxiao Shen, Myle Ott, Michael Auli, and Marc’Aurelio Ranzato. 2019. [Mixture Models for Diverse Machine Translation: Tricks of the Trade](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5719–5728. PMLR.
- Julius Steen and Katja Markert. 2022. [How to Find Strong Summary Coherence Measures? A Toolbox and a Comparative Study for Summary Coherence Measure Evaluation](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6035–6049, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. [Learning to summarize with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.
- Hong Sun and Ming Zhou. 2012. [Joint Learning of a Dual SMT System for Paraphrase Generation](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42, Jeju Island, Korea. Association for Computational Linguistics.
- Teerapaun Tanprasert and David Kauchak. 2021. [Flesch-Kincaid is Not a Text Simplification Evaluation Metric](#). In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. 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)*, pages 90–121, Online. Association for Computational Linguistics.
- Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. [Monolingual and Cross-Lingual Acceptability Judgments with the Italian CoLA corpus](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kees van Deemter. 2024. The Pitfalls of Defining Hallucination. *Computational Linguistics*, 50(2):807–816.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. [Best practices for the human evaluation of automatically generated text](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.
- Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. [Document-level Text Simplification with Coherence Evaluation](#). In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- 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.

- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. **Neural Network Acceptability Judgments**. *Transactions of the Association for Computational Linguistics*, 7:625–641.
- 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.
- Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. **Unsupervised Reference-Free Summary Quality Evaluation via Contrastive Learning**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.
- Hongyi Wu, Xinshu Shen, Man Lan, Shaoguang Mao, Xiaopeng Bai, and Yuanbin Wu. 2023. **A Multi-Task Dataset for Assessing Discourse Coherence in Chinese Essays: Structure, Theme, and Logic Analysis**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6673–6688, Singapore. Association for Computational Linguistics.
- Stratos Xenouleas, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. 2019. **SUM-QE: a BERT-based Summary Quality Estimation Model**. 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 6005–6011, Hong Kong, China. Association for Computational Linguistics.
- 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.
- Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. **SOME: Reference-less Sub-Metrics Optimized for Manual Evaluations of Grammatical Error Correction**. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Udayakumar Nallasamy, and Matthias Paulik. 2019. **Empirical Evaluation of Active Learning Techniques for Neural MT**. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 84–93, Hong Kong, China. Association for Computational Linguistics.
- Min Zhang, Hao Yang, Shimin Tao, Yanqing Zhao, Xiaosong Qiao, Yinlu Li, Chang Su, Minghan Wang, Jiaxin Guo, Yilun Liu, and Ying Qin. 2022. **Incorporating Multilingual Knowledge Distillation into Machine Translation Evaluation**. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers the Digital Economy*, pages 148–160, Singapore. Springer Nature Singapore.
- Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2021. **A Human-machine Collaborative Framework for Evaluating Malevolence in Dialogues**. 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 5612–5623, Online. Association for Computational Linguistics.
- Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. **On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. **Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena**. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Wanzheng Zhu and Suma Bhat. 2020. **GRUEN for Evaluating Linguistic Quality of Generated Text**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.
- Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Trevor Cohn. 2023. **Rethinking Round-Trip Translation for Machine Translation Evaluation**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 319–337, Toronto, Canada. Association for Computational Linguistics.
- Vilém Zouhar, Shehzaad Dhuliawala, Wangchunshu Zhou, Nico Daheim, Tom Kocmi, Yuchen Eleanor Jiang, and Mrinmaya Sachan. 2023. **Poor Man’s Quality Estimation: Predicting Reference-Based MT Metrics Without the Reference**. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1311–1325, Dubrovnik, Croatia. Association for Computational Linguistics.