Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling

Jie Ruan, Xiao Pu, Mingqi Gao, Xiaojun Wan, Yuesheng Zhu

Peking University
{ruanjie,puxiao}@stu.pku.edu.cn, {gaomingqi,wanxiaojun,zhuys}@pku.edu.cn

Abstract

Human evaluation is viewed as a reliable evaluation method for NLG which is expensive and time-consuming. To save labor and costs, researchers usually perform human evaluation on a small subset of data sampled from the whole dataset in practice. However, different selection subsets will lead to different rankings of the systems. To give a more correct inter-system ranking and make the gold standard human evaluation more reliable, we propose a Constrained Active Sampling Framework (CASF) for reliable human judgment. CASF operates through a Learner, a Systematic Sampler and a Constrained Controller to select representative samples for getting a more correct inter-system ranking. Experiment results on 137 real NLG evaluation setups with 44 human evaluation metrics across 16 datasets and 5 NLG tasks demonstrate CASF receives 93.18% top-ranked system recognition accuracy and ranks first or ranks second on 90.91% of the human metrics with 0.83 overall inter-system ranking Kendall correlation. Code and data are publicly available online.

Introduction

Evaluation of NLG systems remains challenging. The reason is that similar content in text can often be expressed in various ways, and the same output of the NLG system may need to satisfy multiple goals in different aspects (2020; 2022). Hence, reliable automatic metrics are complex to design (2017; 2009). Human evaluation is generally considered to be a more reliable evaluation way in natural language generation tasks (2020; 2018; 2015). However, human judgment is viewed as expensive, time-consuming, and lacks standardized evaluation procedures (2020; 2020; 2022).

To save labor and costs, human evaluation is usually performed on a small subset sampled from the dataset in practice. Researchers compare the average scores of the systems on this subset to obtain a ranking between the systems. However, different sample subsets will lead to different rankings of the systems. We re-evaluated 137 real NLG evaluation setups on 44 human metrics across 16 datasets and 5 NLG tasks. Results show that 87.5% of datasets have different inter-system rankings across 5 times of random sampling. Since research is driven by evaluation, focusing on the final ranking of systems, it is vital to design a more reliable evaluation method to obtain the correct inter-system ranking.

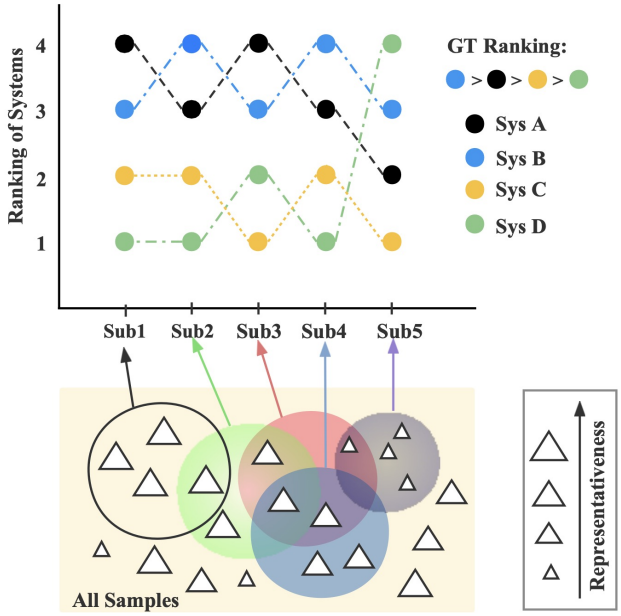

Figure 1: Conducting human evaluations on different sample subsets (Sub) can obtain different inter-system rankings. The lower part shows the same sampling method obtains different subsets at different sampling times. The upper part shows the ranking obtained from the corresponding subsets. "Sys" represents system and "GT" represents Ground Truth.

We randomly select 1404 papers from ACL, EMNLP and COLING in the past 2 years and find that 270 papers select a subset of the dataset for manual evaluation to save labor and cost (details are in the Survey section of the Appendix). The survey results show that random sampling is the most vital sampling method, accounting for 60.7%, and the rest 39.3% of the papers do not mention the sampling method they used. Random sampling is widely used in human evaluation sampling for its simplicity. However, random sampling can be risky (2022). On the one hand, random sampling can lead to clustered selection, a phenomenon in which randomly se-

lected samples are uncommonly close together in a population (as shown in the black and purple circle in Figure 1). On the other hand, random sampling may have the risk of data manipulation. Researchers can choose samples at will or conduct multiple random sampling to select a favorite subset, which will lead to unfair evaluation results. Since different sampling subsets may result in different inter-system rankings in human judgment, it is difficult to reliably select the best system. We urgently need a better sampling method to deliver reliable human evaluation with low labor and cost.

In this paper, we focus on improving the reliability of the gold standard human evaluation with limited cost and time used for human annotation. Specifically, we explore the problem of clustered selection and data manipulation for manual evaluation sampling and propose a Constrained Active Sampling Framework (CASF) for reliable human judgment. The proposed CASF consists of a Learner, a Systematic Sampler and a Constrained Controller. CASF obtains a representative subset of samples in multiple sampling phases. In each sampling phase, the Learner predicts the quality score for samples and feeds the quality score of each sample to the Systematic Sampler. Then, the Systematic Sampler and the Constrained Controller work together to select representative samples with lower redundancy for the sampling phase. Samples collected in each phase are not duplicates of those collected in previous phases, and will be directly subjected to human evaluation, and the newly labeled ones will also be used to update the Learner.

The main contributions are as follows: 1) We investigate and experimentally analyze the sampling problem for the gold standard human evaluation in natural language generation. 2) We propose a Constrained Active Sampling Framework (CASF) for the sampling problem in manual evaluation. The proposed CASF can solve the problem of clustered selection and data manipulation for human evaluation sampling. 3) We re-evaluate 137 real NLG evaluation setups on 44 human evaluation metrics across 16 datasets and 5 NLG tasks. Experiment results demonstrate the proposed method ranks first or ranks second on 90.91% of the human metrics and receives 93.18% top-ranked system recognition accuracy. To ease the adoption of reliable sampling, we release a constrained active sampling tool. We strongly recommend using CASF to sample test instances for human evaluation. Our tool, code and data are publicly available online.[1]

Methodology

Problem Statement

The goal of sampling in human evaluation is to select a subset with the intention of estimating the inter-system ranking of the whole sample population. Ideally, the obtained subset should cover more representative samples of the population. A good sampling method will result in a more correct inter-system ranking calculated through the sampling subset.

The general evaluation sampling problem is as follows. Given a data set $D = \{(x_i, \mathcal{Y}_i, \mathcal{Q}_i)\}_{i=1}^{N}$ where N is the size of the whole sample population, x_i represents a data input, $\mathcal{Y}_i$ is the corresponding set of generated outputs, $\mathcal{Q}_i$ is the corresponding set of human score vectors. The generated output set $\mathcal{Y}_i$ consists of M system outputs and is denoted as $\mathcal{Y}_i = \{y_{i1}, ..., y_{ij}\}_{j=1}^{M}$, where y_{ij} represents the j-th system generated output of the i-th sample. The human score vector set $\mathcal{Q}_i$ consists of the corresponding human score vector for each system output and is denoted as $\mathcal{Q}_i = \{\mathbf{q}_{i1}, ..., \mathbf{q}_{ij}\}_{j=1}^{M}$. Since human evaluation is usually carried out in multiple aspects, we use a vector to represent human evaluation results from multiple aspects for each system. Each human score vector $\mathbf{q}_{ij}$ consists of K human annotation metrics from different aspects and is denoted as $\mathbf{q}_{ij} = (q_{ij1}, ..., q_{ijK})$. Eventually there will be separate inter-system ranking on each aspect. Let $\mathcal{H}$ represent the final sample subset. Function ψ calculates the mean scores of each system in the sample set for each human evaluation aspect and gives the ranking among systems. $\mathcal{P}$ calculates the similarity between two inter-system rankings. The overall objective of sampling and constraint is as follows:

$$\begin{aligned} \text{minimize} \quad & -\mathcal{P}[\psi(\mathcal{H}), \psi(D)], \\ \text{subject to} \quad & |\mathcal{H}| = r \times N, \end{aligned}$$

where r is the sampling rate, $|.|$ refers to the cardinality of a sample set and ψ first calculates the average human scores in each aspect of each system in the sample set, and then gives

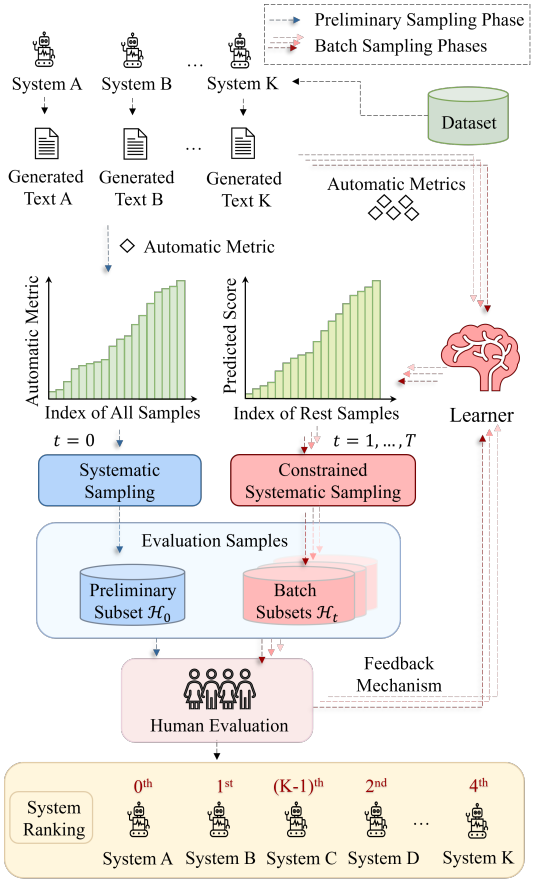

Figure 2: Constrained Active Sampling Framework

[1]https://github.com/EnablerRx/CASF

the inter-system ranking of each human indicator according to the mean score of each system.

Sample Representativeness

Taking representative samples allows for a more complete evaluation of the overall performance of the system. Inspired by the theoretical model of summarization (Peyrard 2019), the *Representativeness* of samples can be measured in two aspects, including *Quality Diversity* and *Redundancy*. *Quality Diversity* represents the diversity of sample quality levels, that is, the sampled subset should contain samples of various quality levels. Evaluation on qualitatively diverse subsets of samples allows the system to better reflect the performance of all samples. Quality is the average quality of generated outputs of the sample. More comprehensive coverage of samples of different qualities will result in a better *Quality Diversity*. *Redundancy* indicates the degree of similarity or duplication among the generated outputs of samples.

Constrained Active Sampling Framework

Overall Framework The proposed Constrained Active Sampling Framework aims to select representative samples for human evaluation in multiple phases to get a more correct inter-system ranking. The proposed CASF operates through a Learner, a Systematic Sampler and a Constrained Controller. The goal of the Learner is to predict the quality of samples and give a ranking of sample quality by a regressor. The Systematic Sampler divides samples into multiple buckets according to the sample quality ranking given by the Learner. The Constrained Controller controls the *Redundancy* of samples and selects a final sample from each bucket given by the Systematic Sampler.

The proposed Constrained Active Sampling Framework is shown in Figure 2. There are several sampling phases denoted by t, an preliminary sampling phase $t = 0$ (the left branch in Figure 2) and T batch active sampling phases $t = 1, ..., T$ (the right branch in Figure 2). In the preliminary sampling phase, alternate quality scores for all samples are calculated through an automated metric, as the Learner is not ready to use yet. The Systematic Sampler, then, selects a small preliminary subset of samples $\mathcal{H}_0$ as part of the final sample subset $\mathcal{H}$ according to the given quality ranking. The selected samples are then evaluated by human beings. In the current batch active sampling phase, samples selected in all previous phases together with the corresponding human scores, then, are fed to the regressor of the learner, and the regressor of the learner is updated and applied to predict the quality of the rest samples with the sample's scores over various automatic metrics as features. After that, the Systematic Sampler and Constrained Controller work together to choose batch subset $\mathcal{H}_t$ from the rest samples for the t-th batch active sampling phase as part of the final samples. Then, the samples selected in the t-th phase are subjected to human evaluation for use in the subsequent sampling phases. The final sample set $\mathcal{H}$ consists of batch subsets from each phase $\mathcal{H}_t$. We conduct experiments to explore the determination of the number of phases and the sampling ratio of each phase in the Phases and Associated Sampling Ratios section.

Learner and Sample Quality Estimating the quality of the samples is a vital step in CASF. Since the quality of samples is difficult to define and calculate directly, we propose a Learner to predict the human scores as the quality scores for the rest samples for selection at each phase t (except the preliminary phase). As various automatic metrics can measure the characteristics of samples in different aspects and are easy to calculate with lower cost, we use scores of automatic metrics as features to predict the quality of samples.

Note that in the preliminary phase, the quality of samples is simply estimated by an automatic metric. In each of the batch active sampling phases, the Learner receives feedback from human annotators and update its parameters. After that, it utilizes the scores of automatic metrics to predict the quality score for each sample. The Learner will then provide the quality ranking $\{p_t(i)\}_{i=1}^{N-|\mathcal{H}|}$ of samples at each batch t, where i is the sample index and the number of the rest samples for selection in each phase is $N - |\mathcal{H}|$.

The main objective of the Learner g is to map x_i to the corresponding human score vector set $\mathcal{Q}_i$. Since there are multiple elements in $\mathcal{Q}_i$, we standardize scores for each human evaluation aspect and use the sum of each element in $\mathcal{Q}_i$, which is the sum of human scores for all aspects of all NLG systems under sample x_i, to represent $\mathcal{Q}_i$. The objective is to minimize the following loss function:

$$\underset{\theta_t}{\operatorname{argmin}} \sum_{i=1}^{|\mathcal{H}|} \mathcal{L}\left(g(x_i;\theta_t), \sum_{j=1}^{M}\sum_{k=1}^{K} q_{ijk}\right),$$

where $|\mathcal{H}|$ is the number of samples selected in the final subset and θ_t is the parameter of Learner g in the t-th phase. The predicted quality scores $\{s_t(i)\}_{i=1}^{N-|\mathcal{H}|}$ for the rest samples at each phase t are calculated as follows:

$$\{s_t(i)\}_{i=1}^{N-|\mathcal{H}|} = \{g(x_i;\theta_t)\}_{i=1}^{N-|\mathcal{H}|}.$$

Specifically, the Learner first calculates the results of each automatic metric based on the output of each NLG system from the input sample. Then, the automatic metric results under each NLG system will be fed as features into the Learner's regressor. Eight popular NLG metrics are chosen as the automatic metrics set (details are in the Automatic Metric for Preliminary Phase section) of CASF. Due to the small number of samples and features mainly containing automatic metrics' scores, we explore several popular learning methods and recommend choosing Gradient Boosting Decision Tree (GBDT) (Friedman 2001) as the regressor of the Learner. Full experimental results are in the Learner Selection section of Appendix. The loss function is the least squares method (2007), which is commonly used in GBDT.

Systematic Sampler Systematic sampling has advantage of eliminating clustered selection problem and can reduce the risk of favoritism, which meets our motivation. Therefore, we adopt the systematic sampling method (Yates 1948) sorted by relevant signs as the sampling core of CASF. The Systematic Sampler selects representative initial samples and candidate samples according to the quality ranking of samples. Specifically, the Systematic Sampler first divides

the $N_t = N - |\mathcal{H}|$ samples for the t-th phase into n_t buckets according to the given quality ranking $\{p_t(i)\}_{i=1}^{N-|\mathcal{H}|}$. n_t is the number of samples to be selected at the t-th phase. Samples with quality ranking $p_t \in [e \times \lfloor \frac{N_t}{n_t} \rfloor, (e+1) \times \lfloor \frac{N_t}{n_t} \rfloor)$ are divided into the same bucket, where $e = 0, 1, ..., n_t$. The samples with quality rank $p_t = e \times \lfloor \frac{N_t}{n_t} \rfloor$ are selected as the **initial selection samples**. And the rest samples in each bucket are **candidate samples**.

Constrained Controller The proposed Constrained Controller controls the *Redundancy* of samples and selects one sample from each of the buckets divided by the Systematic Sampler to form a final sample subset (as shown in Figure 3). Since the Systematic Sampler selects initial samples at a regular interval, which makes the distribution of the initial subset align closely with the overall distribution, we aim to preserve the original sampling intervals as much as possible while controlling the Redundancy to maintain the representativeness of the sample subset.

Specifically, we define objective function Obj as the quality ranking distance between the current sample x_i and the initial selection sample in each bucket. We also define violation function Vio to calculate the *Redundancy* between the current sample x_i and the final samples. Since the bi-gram similarity (Kondrak 2005) is regarded as a simple and effective method to calculate the redundancy between texts, we calculate the *Redundancy* by calculating the bi-gram similarity between the outputs generated for the sample and that for the final samples. A sample x_i is called feasible if $\mathrm{Vio}(x_i) = 0$, which means it is not redundant with the selected final samples. Otherwise, x_i is infeasible.

The Constrained Controller is summarized into 3 rules:

$$\begin{cases} rule\ 1: & x_i \prec x_j, \quad \text{if} \quad x_i \text{ is infeasible, } x_j \text{ is feasible;} \\ rule\ 2: & x_i \prec x_j, \quad \text{if} \quad \mathrm{Vio}(x_i) > \mathrm{Vio}(x_j), \\ & \qquad\qquad\ x_i \text{ is infeasible, } x_j \text{ is infeasible;} \\ rule\ 3: & x_i \prec x_j, \quad \text{if} \quad \mathrm{Obj}(x_i) > \mathrm{Obj}(x_j), \\ & \qquad\qquad\ x_i \text{ is feasible, } x_j \text{ is feasible,} \end{cases}$$

where x_i is the i-th sample and $x_i \prec x_j$ means x_j is a better choice. $rule\ 1$ means the Constrained Controller tends to select samples that are not redundant. $rule\ 2$ represents that if two samples are both redundant with the final samples, the Constrained Controller tends to select samples with less redundancy. $rule\ 3$ demonstrates that if two samples are both not redundant with the final samples, the Constrained Controller tends to select samples with ranks as close as possible to those of the initial selection samples.

In Figure 3, the rest samples for selection are first re-indexed, and then re-ordered according to Learner's predicted quality score. The system sampler divides samples into three buckets based on quality ranking and marks initial selection sample for each bucket. In the first bucket, only sample 3 is feasible, that is, sample 3 is not redundant with existing final samples. Thus, Sample 3 is selected as the final sample according to $rule\ 1$. In the second bucket, none of the three samples is feasible, so sample 0 with the smallest redundancy is selected as the final sample according to $rule\ 2$. In the third bucket, all samples are feasible, and sample 5 is the initial selection sample and it is selected by default or according to $rule\ 3$.

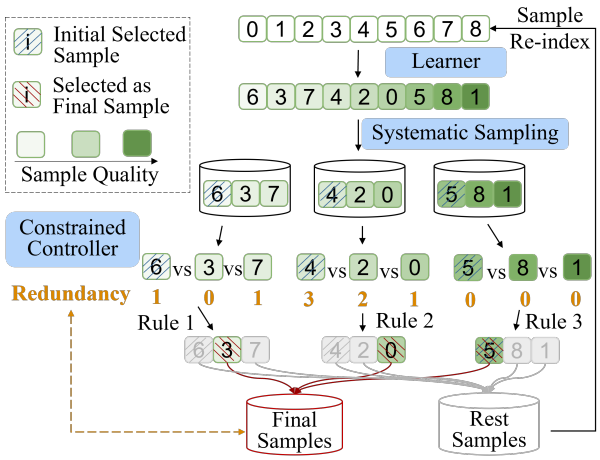

Figure 3: Example of systematic sampler and constrained controller cooperating to select final samples

Experimental Setup

Tasks and Datasets

We conduct experiments on 44 human metrics across 16 datasets spanning 5 tasks. A total of 137 NLG systems are involved. Details of the datasets, preprocessing and the validation set for hyper-parameters selection are in the Tasks and Dataset section of Appendix. The datasets are: **Summarization (SUM):** We utilize 8 human evaluation datasets of the model generated summarization, which are SummEval (2021), REALSumm (2020), Newsroom (NeR18) (2018), DialSummEval (DialSumm) (2022) and OpenAI-axis1 (OpenAI 1) (2020; 2017), OpenAI-axis2 (OpenAI 2) , OpenAI-CNN/DM1 (OpenAI 3) , and OpenAI-CNN/DM3 (OpenAI 4) . **Machine Translation (MT):** We use 3 datasets collected from WMT news translation tasks (2021) *viz.* newstest2020 en-de (newstest 1), newstest2020 cn-en (newstest 2) and newstest2021 cn-en (newstest 3). **Dialogue Generation (DGen):** We utilize a human annotation dataset of machine-generated dialogues released with the Persona Chat (Persona) (Mehri and Eskenazi 2020) dataset. **Story Generation (SGen):** We use two manual evaluation datasets for story generation namely MANS-ROC (Guan et al. 2021) and MANS-WP (Guan et al. 2021). **Multi-Modal Generation (MMGen):** We use two existing human evaluation datasets namely THUMB-MSCOCO (THUMB) (Kasai et al. 2022) and VATEX-EVAL (VATEX) (Shi et al. 2022).

Evaluation Metric

We select a subset of each dataset and then compute the results for all the human metrics in various aspects. We measure the efficacy of sampling method by computing rankings of candidate models on the subset and their Kendall's Tau correlation (1938) with rankings obtained on the full dataset. We refer to Kendall's treatment (1945) to handle ties.

Dataset	HE Metric	R 1	R 2	R 3	R Mean	H 1	H 2	H 3	H Mean	8M	SM	OL	CASF (ours)
SummEval	coherence	0.85	0.65	0.33	0.61	0.70	0.82	0.92	0.81	0.42	0.42	0.87	**0.95**
	consistency	0.25	0.48	0.43	0.39	0.68	0.02	0.65	0.45	0.30	0.17	**0.53**	**0.53**
	fluency	0.40	0.35	0.52	0.42	0.45	0.45	0.30	0.40	0.35	0.37	**0.52**	0.33
	relevance	0.72	0.60	0.68	0.67	0.65	0.43	0.72	0.60	0.40	0.60	0.45	**0.82**
REALSumm	litepyramid	0.39	0.54	0.44	0.46	0.36	0.38	0.44	0.39	0.33	0.37	**0.54**	**0.54**
NeR18	coherence	1.00	1.00	0.43	0.81	0.90	0.90	0.90	0.90	**1.00**	**1.00**	**1.00**	**1.00**
	fluency	0.52	1.00	1.00	0.84	1.00	0.52	0.90	0.81	**1.00**	**1.00**	**1.00**	**1.00**
	informativeness	1.00	1.00	1.00	**1.00**	0.71	1.00	0.90	0.87	0.71	**1.00**	**1.00**	**1.00**
	relevance	1.00	0.52	1.00	0.84	0.90	0.90	0.90	0.90	**1.00**	**1.00**	**1.00**	**1.00**
DialSumm	consistency	0.74	0.72	0.49	0.65	0.74	0.64	0.62	0.67	0.59	0.56	0.54	**0.77**
	relevance	0.69	0.46	0.64	0.60	0.64	0.69	0.54	0.62	0.23	0.44	0.59	**0.72**
	fluency	0.59	0.56	0.59	0.58	0.38	0.56	0.51	0.49	0.15	0.49	**0.64**	0.62
	coherence	0.67	0.80	0.74	0.74	0.74	0.80	0.59	0.71	0.59	0.67	0.82	**0.90**
OpenAI 1	accuracy	0.80	0.00	1.00	0.60	0.80	1.00	0.80	0.87	0.80	0.00	0.00	**1.00**
	coherence	0.40	0.80	0.00	0.40	0.80	0.20	0.80	0.60	**0.80**	0.40	0.20	**0.80**
	coverage	1.00	1.00	1.00	**1.00**	0.80	0.80	0.80	0.80	0.80	**1.00**	0.80	0.80
	overall	0.80	1.00	1.00	0.93	0.80	1.00	0.80	0.87	0.80	**1.00**	0.80	**1.00**
OpenAI 2	accuracy	0.71	0.43	1.00	0.71	0.62	0.71	0.81	0.71	**1.00**	0.52	0.14	0.90
	coherence	0.24	0.52	0.33	0.37	-0.14	0.24	0.43	0.17	0.24	**0.52**	0.24	0.43
	coverage	1.00	0.71	0.90	0.87	1.00	0.90	1.00	0.97	**1.00**	**1.00**	**1.00**	**1.00**
	overall	0.90	0.71	1.00	0.87	0.62	1.00	0.90	0.84	0.90	0.90	0.90	**1.00**
OpenAI 3	accuracy	0.73	0.82	0.82	0.79	0.87	0.78	0.82	0.82	0.73	0.69	0.78	**0.87**
	coherence	0.51	0.33	0.56	0.47	0.42	0.51	0.56	0.50	0.56	0.20	**0.60**	**0.60**
	coverage	0.38	0.38	0.87	0.54	0.51	0.87	0.51	0.63	**1.00**	**1.00**	0.42	0.87
	overall	0.87	0.51	1.00	0.79	1.00	0.73	0.51	0.75	**1.00**	0.38	0.47	**1.00**
OpenAI 4	accuracy	1.00	0.33	1.00	0.78	1.00	0.33	0.33	0.56	0.33	**1.00**	0.33	**1.00**
	coherence	1.00	1.00	1.00	**1.00**	1.00	1.00	1.00	**1.00**	**1.00**	**1.00**	0.33	**1.00**
	coverage	0.33	1.00	1.00	0.78	0.33	1.00	1.00	0.78	**1.00**	**1.00**	**1.00**	**1.00**
	overall	0.33	1.00	1.00	0.78	0.33	1.00	1.00	0.78	**1.00**	**1.00**	**1.00**	**1.00**
newstest 1	MQM	0.14	0.14	0.14	0.14	0.33	0.14	-0.05	0.14	0.14	**0.33**	0.14	0.14
	pSQM	0.81	0.90	0.90	0.87	0.81	0.90	0.90	0.87	**1.00**	0.90	0.90	**1.00**
newstest 2	MQM	0.79	0.93	0.71	0.81	0.64	0.86	0.71	0.74	0.14	**0.93**	0.86	**0.93**
	pSQM	0.43	0.36	0.79	0.52	0.29	0.86	0.43	0.52	0.36	**0.93**	0.79	0.79
newstest 3	MQM	0.00	-0.13	-0.05	-0.06	-0.05	-0.03	-0.05	-0.04	**0.46**	0.13	0.00	0.03
Persona	Understandable	0.33	-1.00	0.33	-0.11	-1.00	0.33	0.33	-0.11	**0.33**	**0.33**	**0.33**	**0.33**
	Natural	0.33	-1.00	1.00	0.11	1.00	-1.00	0.33	0.11	0.33	0.33	0.33	**1.00**
	Maintains Context	1.00	1.00	1.00	**1.00**	1.00	1.00	1.00	**1.00**	-1.00	**1.00**	**1.00**	**1.00**
	Interesting	1.00	1.00	1.00	**1.00**	1.00	0.33	1.00	0.78	**1.00**	**1.00**	**1.00**	**1.00**
	Uses Knowledge	1.00	1.00	1.00	**1.00**	-1.00	1.00	1.00	0.33	**1.00**	**1.00**	**1.00**	**1.00**
	Overall Quality	1.00	1.00	1.00	**1.00**	1.00	1.00	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
MANS-ROC	overall	1.00	1.00	1.00	**1.00**	1.00	1.00	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
MANS-WP	overall	1.00	0.80	0.80	0.87	0.80	1.00	1.00	0.93	**1.00**	**1.00**	**1.00**	**1.00**
THUMB	overall	1.00	0.80	1.00	0.93	1.00	1.00	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
VATEX	consistency	0.60	1.00	0.60	0.73	0.60	1.00	1.00	0.87	**1.00**	**1.00**	**1.00**	**1.00**
Overall Performance		0.69	0.61	0.75	0.68	0.61	0.67	0.65	0.72	0.67	0.72	0.68	**0.83**

Table 1: Kendall's Tau of methods on 16 datasets across 5 NLG tasks. 'HE Metric' indicates different human evaluation aspects in a dataset. Bold number indicates that the method has the best performance among all methods under the corresponding aspect. Underlined number indicates the method ranks second.

Method	SUM	MT	DGen	SGen	MMGen	Overall
R	0.76	0.87	0.78	0.67	1.00	0.76
H	0.80	0.67	0.78	0.67	1.00	0.78
8M	0.83	0.80	0.83	1.00	1.00	0.84
SM	0.90	1.00	0.83	1.00	1.00	0.91
OL	0.69	0.80	1.00	1.00	1.00	0.77
CASF	0.93	0.80	1.00	1.00	1.00	**0.93**

Table 2: Top-ranked accuracy on 5 NLG tasks. 'Overall' shows the average result on all human metrics from all tasks.

Comparison of Methods

The comparison methods are selected based on the survey of evaluation sampling methods in 1404 papers where Random and Heuristic are the main sampling methods for NLG human evaluation. We also include some ablation methods. The comparison methods are: **Random Sampling (R)** randomly sample the dataset and is performed 3 times (2008; 2022; 2007) to reflect real sampling scenarios. Results of each time and the average result are recorded. **Heuristic Sampling (H)** (2022) first sorts the samples according to the average length of the generated sentences. Then, Heuristic randomly collects a small number of samples with extreme sentence length and a large number of samples with normal sentence length. Heuristic is performed 3 times. **Eight Metric (8M):** CASF with only the preliminary sampling phase which normalizes the score obtained by the 8 automatic metrics used in CASF and calculates the average score. **Single Metric (SM):** CASF with only the preliminary sampling phase which uses the automatic metric used in the preliminary sampling phases of CASF. **Online Sampling (OL):** CASF without Constrained Controller. We compare methods with 50% sampling rate. Results for other sampling ratios are in Different Sampling Ratio section of Appendix. In addition, the number of phases and the sampling ratio of each phase are 5 and 10%. The determination of these parameters is shown in the Phases and Associated Sampling Ratios section. We also treat the sample size as an independent variable and results are shown in the Appendix.

Results and Analysis

Comparison Results

Full Inter-System Ranking Accuracy According to results on validation set (Automatic Metrics for Preliminary Sampling Phase section of Appendix), We select MOVERSCORE (Zhao et al. 2019) for calculating sample quality in the preliminary sampling phase. Inter-system ranking accuracy of methods on 16 datasets across 5 NLG tasks are shown in Table 1. The results show Random have large fluctuations. For example, in the newstest2020 cn-en dataset of MT task, different times of random sampling result in different inter-system correlation. This shows the risky of widely using Random in evaluation. CASF ranked first on 79.55% of human metrics and ranked first or ranked second on 90.91% of metrics. This shows CASF can better select representative samples to get a more accurate ranking. Results of the remaining human metrics, although not ranking first, are still acceptable and close to the best results. These acceptable results appear as we measure the quality of each sample in the dataset. However, human evaluation in different aspects is conducted in the same dataset. The overall scores can represent the overall evaluation results. We use Wilcoxon signed ranks (2006) to test the results of Random and Heuristic (both iterated 10000 times) with CASF in 44 human metrics. Results show CASF is statistically outperforming Random, Heuristic and other methods with $p = 0.00010$, $p = 0.00009$ and $p < 0.05$.

	Sys A	Sys B	Sys C	Sys D	Sys E	Tau
GT	3	4	1	0	2	
R 1	4	3	1	0	2	0.80
R 2	3	1	4	0	2	0.00
R 3	1	3	4	0	2	0.20
H 1	1	4	3	0	2	0.40
H 2	1	3	4	0	2	0.20
H 3	4	3	1	0	2	0.80
8M	4	3	1	0	2	0.80
SM	3	1	4	0	2	0.00
OL	3	1	4	0	2	0.00
CASF	3	4	1	0	2	**1.00**

Figure 4: Inter-system ranking of human evaluation aspect 'accuracy' of OpenAI 1. "GT" is the inter-system ranking on the entire dataset. Sampling rate is 50%. "Sys" represents system. Rankings in red indicate incorrect rankings.

Top-Ranked System Accuracy One of the important goals of evaluation is to select the top-ranked system. Accurately selecting the best system with limited manpower can help the NLG field to keep good systems and eliminate poor ones. Thus, we explore the ability of CASF to identify the top-ranked system. As shown in Table 2, CASF achieves 93.18% top-ranked system recognition accuracy in 44 human evaluation metrics involving 137 NLG systems. For typical NLG tasks like DGen, SGen and MMGen, CASF achieves 100% identification accuracy. Experimental results also showed CASF was statistically outperforming the popular Random and Heuristic at the $p < 0.05$ level.

Case Study Taking the human aspect accuracy in the OpenAI 1 (Stiennon et al. 2020; Völske et al. 2017) dataset as an example, CASF obtains an accurate inter-system ranking as shown in Figure 4. The 3 times of random sampling obtained different inter-system rankings, and the ranking of the first system fluctuated between the first and fourth, with great volatility. This confirms the problem we raised about the risk of random sampling, making evaluation unreliable. CASF selects the same subset in multiple times, and the variance of the inter-ranking accuracy obtained by multiple sampling times is 0 (Learner Selection section of Appendix). Since CASF selects representative samples, it obtains more accurate inter-system rankings, making evaluation more reliable.

M	#P	P-R	B-R	Tau	M	#P	P-R	B-R	Tau	M	#P	P-R	B-R	Tau	M	#P	P-R	B-R	Tau
	2	0.25	0.25	0.75		2	0.10	0.40	0.73		2	0.05	0.45	0.74		2	0.15	0.35	0.73
	3	0.17	0.17	0.76		3	0.10	0.20	0.75		3	0.05	0.23	0.74		3	0.15	0.18	0.77
	4	0.13	0.13	0.76		4	0.10	0.13	0.80		4	0.05	0.15	0.77		4	0.15	0.12	0.76
	5	0.10	0.10	**0.83**		**5**	0.10	0.10	**0.83**		**5**	0.05	0.11	**0.77**		5	0.15	0.09	0.76
A	6	0.08	0.08	0.72	F	6	0.10	0.08	0.75	F	6	0.05	0.09	0.73	F	6	0.15	0.07	0.71
	7	0.07	0.07	0.72		7	0.10	0.07	0.69		7	0.05	0.08	0.69		7	0.15	0.06	0.73
	8	0.06	0.06	0.70		8	0.10	0.06	0.73		8	0.05	0.06	0.72		**8**	0.15	0.05	**0.79**
	9	0.06	0.06	0.73		9	0.10	0.05	0.72		9	0.05	0.06	0.72		9	0.15	0.04	0.75
	10	0.05	0.05	0.75		10	0.10	0.04	0.73		10	0.05	0.05	0.75		10	0.15	0.04	0.70

Table 3: Experimental results on 44 human metrics with different mode (M) (Average (A) and Preliminary-Fixed (F)), number of phases (#P), preliminary sample ratio (P-R) and batch sampling ratio (B-R) of each phase for the proposed CASF.

Automatic Metric for Preliminary Phase

We choose automatic metrics commonly used in NLG as our automatic metrics set, including BERT-SCORE (2019), MOVER-SCORE (2019), ROUGE-1 (2004), ROUGE-2, ROUGE-L, BART-SCORE (2021), BLEU (2002) and METEOR (2005). We apply each metric to calculate sample quality in the preliminary sampling phase of CASF in Table 4. Results show sample quality calculated on MOVERSCORE get a more correct ranking. This shows the ability to calculate sample quality of contextual-embedding-based metric MOVER-SCORE. Traditional metric METEOR ranks second. Full results are in Appendix.

Phases and Associated Sampling Ratios

We conduct experiments to explore the influence of the number of phases and the sampling ratio of each phase for CASF. Results at the sampling rate of 50% on 16 datasets are shown in Table 3. In average mode, all phases are sampled in equal proportions. In the preliminary-fixed mode, we fix the preliminary sampling ratio, and the batch sampling ratio is divided equally according to the number of iteration phases and the total sampling ratio. Results show that performance is better when the number of iteration phases is 5 in most cases. It is simple and effective to sample each phase according to the total sampling rate and the number of phases.

Significant Information Retention Accuracy

Previous work (2022) focused on identifying top-ranked systems, and we further explored giving more accurate overall inter-system rankings and tested the significant information retention accuracy on sample subsets, that is, to test whether the subset can preserve the significance of ranking among systems. Results showed CASF outperforms Random and Heuristic. Details are in the Appendix.

Metric	SUM	MT	DGen	SGen	MMGen	Avg
BERT-S	0.74	0.58	0.67	1.00	1.00	0.73
MOVER-S	0.84	0.58	0.89	1.00	1.00	**0.83**
ROUGE-1	0.73	0.57	0.67	0.30	1.00	0.70
ROUGE-2	0.73	0.55	0.56	1.00	0.80	0.70
ROUGE-L	0.72	0.52	0.89	1.00	1.00	0.75
BART-S	0.60	0.44	0.89	0.90	0.80	0.64
BLEU	0.72	0.37	0.56	1.00	0.80	0.67
METEOR	0.78	0.54	0.89	1.00	1.00	<u>0.79</u>

Table 4: Results of CASF pre-ranking on different automatic metrics. "-S" indicates "-Score". "Avg" represents the average result on all human metrics from all tasks.

Related Work

Previous works (2014; 2015; 2014; 2016) adopt TrueSkill (2006) to rank NLG methods with pairwise human evaluation. Sakaguchi and Van Durme (2018) introduce a method for system quality estimation from pairwise annotation by human judgment. Hashimoto et al. (2019) propose an evaluation mechanism to calculate a model's sampling probabilities. Chaganty, Mussman, and Liang (2018) utilize control variates to obtain an unbiased estimator with lower cost than only using human evaluation. Mendonça et al. (2021) adopt online learning to find the best systems for machine translation. Wei et al. (2022) study the power on pairwise direct assessment comparisons. A recent work (2022) introduces Active Evaluation to identify the top-ranked system with less pairwise human annotations. There is still a vacancy in the research to derive a complete inter-system ranking based on the results of direct human scoring for general NLG tasks. Yates (1948) proposed Systematic Sampling. ILDAE (2022) calculates the difficulty score of the sample and uses a simple sampling method for Natural Language Inference. However, ILDAE is not suitable for NLG since there is no direct confidence value in NLG methods. To the best of our knowledge, this paper is the first work to extensively study the sampling method for direct scoring to get the whole inter-system ranking in NLG human evaluation.

Conclusion

In this paper, we focused on giving a more correct inter-system ranking for reliable human evaluation with limited time and cost. We propose CASF and show the overall inter-system Kendall correlation improved by 41% to 0.83 compared to the widely used random sampling in 44 human evaluation metrics across 16 datasets in 5 NLG tasks. CASF ranked first or ranked second among all comparison methods on up to 90.91% of the human metrics. We release a tool and we strongly recommend using CASF for reliable human evaluation to get a more reliable inter-system ranking.

Acknowledgements

This work was supported by National Key R&D Program of China (2021YFF0901502), National Science Foundation of China (No. 62161160339), State Key Laboratory of Media Convergence Production Technology and Systems and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Abdi, H.; et al. 2007. The method of least squares. *Encyclopedia of measurement and statistics*, 1: 530–532.

Banerjee, S.; and Lavie, A. 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*, 65–72.

Bethard, S. 2022. We need to talk about random seeds. *arXiv preprint arXiv:2210.13393*.

Bhandari, M.; Gour, P.; Ashfaq, A.; Liu, P.; and Neubig, G. 2020. Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100*.

Bhatnagar, R.; Ganesh, A.; and Kann, K. 2022. CHIA: CHoosing Instances to Annotate for Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 7299–7315.

Bojar, O.; Buck, C.; Federmann, C.; Haddow, B.; Koehn, P.; Leveling, J.; Monz, C.; Pecina, P.; Post, M.; Saint-Amand, H.; et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, 12–58.

Bojar, O.; Chatterjee, R.; Federmann, C.; Haddow, B.; Huck, M.; Hokamp, C.; Koehn, P.; Logacheva, V.; Monz, C.; Negri, M.; Post, M.; Scarton, C.; Specia, L.; and Turchi, M. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 1–46. Lisbon, Portugal: Association for Computational Linguistics.

Celikyilmaz, A.; Clark, E.; and Gao, J. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Chaganty, A. T.; Mussman, S.; and Liang, P. 2018. The price of debiasing automatic metrics in natural language evaluation. *arXiv preprint arXiv:1807.02202*.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7: 1–30.

Fabbri, A. R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; and Radev, D. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9: 391–409.

Freitag, M.; Foster, G.; Grangier, D.; Ratnakar, V.; Tan, Q.; and Macherey, W. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. arXiv:2104.14478.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Gao, M.; and Wan, X. 2022. DialSummEval: Revisiting Summarization Evaluation for Dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5693–5709. Seattle, United States: Association for Computational Linguistics.

Gatt, A.; and Krahmer, E. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61: 65–170.

Gkatzia, D.; and Mahamood, S. 2015. A snapshot of NLG evaluation practices 2005-2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, 57–60.

Grusky, M.; Naaman, M.; and Artzi, Y. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Guan, J.; Zhang, Z.; Feng, Z.; Liu, Z.; Ding, W.; Mao, X.; Fan, C.; and Huang, M. 2021. OpenMEVA: A Benchmark for Evaluating Open-ended Story Generation Metrics. arXiv:2105.08920.

Hashimoto, T. B.; Zhang, H.; and Liang, P. 2019. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*.

Herbrich, R.; Minka, T.; and Graepel, T. 2006. TrueSkill™: a Bayesian skill rating system. *Advances in neural information processing systems*, 19.

Howcroft, D. M.; Belz, A.; Clinciu, M.-A.; Gkatzia, D.; Hasan, S. A.; Mahamood, S.; Mille, S.; Van Miltenburg, E.; Santhanam, S.; and Rieser, V. 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*, 169–182.

Kasai, J.; Sakaguchi, K.; Dunagan, L.; Morrison, J.; Bras, R. L.; Choi, Y.; and Smith, N. A. 2022. Transparent Human Evaluation for Image Captioning. In *Proc. of NAACL*.

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

Kendall, M. G. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3): 239–251.

Kondrak, G. 2005. N-gram similarity and distance. In *International symposium on string processing and information retrieval*, 115–126. Springer.

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

Mehri, S.; and Eskenazi, M. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

Mendonça, V.; Rei, R.; Coheur, L.; Sardinha, A.; and Santos, A. L. 2021. Online learning meets machine translation evaluation: Finding the best systems with the least human effort. *arXiv preprint arXiv:2105.13385*.

Mohankumar, A. K.; and Khapra, M. M. 2022. Active Evaluation: Efficient NLG Evaluation with Few Pairwise Comparisons. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8761–8781.

Novikova, J.; Dušek, O.; Curry, A. C.; and Rieser, V. 2017. Why we need new evaluation metrics for NLG. *arXiv preprint arXiv:1707.06875*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Peyrard, M. 2019. A Simple Theoretical Model of Importance for Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1059–1073.

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

Sakaguchi, K.; Napoles, C.; Post, M.; and Tetreault, J. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4: 169–182.

Sakaguchi, K.; Post, M.; and Van Durme, B. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 1–11.

Sakaguchi, K.; and Van Durme, B. 2018. Efficient online scalar annotation with bounded support. *arXiv preprint arXiv:1806.01170*.

Shi, Y.; Yang, X.; Xu, H.; Yuan, C.; Li, B.; Hu, W.; and Zha, Z. 2022. EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Varshney, N.; Mishra, S.; and Baral, C. 2022. ILDAE: Instance-Level Difficulty Analysis of Evaluation Data. *arXiv preprint arXiv:2203.03073*.

Völske, M.; Potthast, M.; Syed, S.; and Stein, B. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, 59–63.

Wan, X.; and Xiao, J. 2008. CollabRank: towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 969–976.

Wan, X.; and Yang, J. 2007. CollabSum: exploiting multiple document clustering for collaborative single document summarizations. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 143–150.

Wei, J. T.-Z.; Kocmi, T.; and Federmann, C. 2022. Searching for a higher power in the human evaluation of MT. *arXiv preprint arXiv:2210.11612*.

Yates, F. 1948. Systematic sampling. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 241(834): 345–377.

Yuan, W.; Neubig, G.; and Liu, P. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34: 27263–27277.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhao, W.; Peyrard, M.; Liu, F.; Gao, Y.; Meyer, C. M.; and Eger, S. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

Zhou, K.; Blodgett, S. L.; Trischler, A.; Daumé III, H.; Suleman, K.; and Olteanu, A. 2022. Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications. *arXiv preprint arXiv:2205.06828*.