DecompEval: Evaluating Generated Texts as Unsupervised Decomposed Question Answering

Pei Ke[1], **Fei Huang**[1], **Fei Mi**[2], **Yasheng Wang**[2], **Qun Liu**[2], **Xiaoyan Zhu**[1], **Minlie Huang**[1*]

[1]The CoAI Group, DCST, Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
[2]Huawei Noah's Ark Lab, China
kepei1106@outlook.com, f-huang18@mails.tsinghua.edu.cn
{mifei2, wangyasheng, qun.liu}@huawei.com, {zxy-dcs, aihuang}@tsinghua.edu.cn

Abstract

Existing evaluation metrics for natural language generation (NLG) tasks face the challenges on generalization ability and interpretability. Specifically, most of the well-performed metrics are required to train on evaluation datasets of specific NLG tasks and evaluation dimensions, which may cause over-fitting to task-specific datasets. Furthermore, existing metrics only provide an evaluation score for each dimension without revealing the evidence to interpret how this score is obtained. To deal with these challenges, we propose a simple yet effective metric called DecompEval. This metric formulates NLG evaluation as an instruction-style question answering task and utilizes instruction-tuned pre-trained language models (PLMs) without training on evaluation datasets, aiming to enhance the generalization ability. To make the evaluation process more interpretable, we decompose our devised instruction-style question about the quality of generated texts into the subquestions that measure the quality of each sentence. The subquestions with their answers generated by PLMs are then recomposed as evidence to obtain the evaluation result. Experimental results show that DecompEval achieves state-of-the-art performance in untrained metrics for evaluating text summarization and dialogue generation, which also exhibits strong dimension-level / task-level generalization ability and interpretability[1].

1 Introduction

Recently, pre-trained language models (PLMs) such as GPT (Brown et al., 2020), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020) have achieved promising performance in natural language generation (NLG) tasks, such as text summarization (Zhang et al., 2020a) and dialogue generation (Zhang et al., 2020c). As the quality of generated texts gradually approaches that of human-written texts, there is an increasing demand for automatic evaluation metrics of generated texts.

However, existing evaluation metrics are still struggling to measure the quality of generated texts accurately. Traditional metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004) rely on n-gram overlap between generated texts and reference texts, which fail to detect the issues in the content of generated texts (Gehrmann et al., 2022). Recent works resort to model-based evaluation metrics to compute the similarity between generated texts and reference texts based on contextual representations from pre-trained models (Zhao et al., 2019; Zhang et al., 2020b) or adopt the score of language modeling (Yuan et al., 2021) / masked language modeling (Ke et al., 2022; Colombo et al., 2022) for evaluation. Other works choose to train evaluation models on the evaluation datasets to fit human scores (Shen et al., 2017; Sellam et al., 2020) or distinguish human-written texts from negative samples (Guan and Huang, 2020; Zhong et al., 2022), aiming to obtain higher correlations with human judgments in various evaluation dimensions (such as coherence and consistency) of specific datasets.

We argue that there are two main challenges in building an evaluation metric for text generation: 1) **Generalization Ability**: Most of the existing metrics that have high correlations with human judgments on evaluation datasets are directly trained on the corresponding datasets (Sellam et al., 2020; Guan and Huang, 2020; Zhong et al., 2022). This may result in over-fitting to task-specific data and harm their generalization ability to other NLG tasks and dimensions (Ke et al., 2022). 2) **Interpretability**: Although recently proposed evaluation metrics can measure the quality of generated texts from multiple dimensions, they only provide an evaluation score for each dimension without giving evidence to interpret how they predict this score

[*] Corresponding author

[1]The codes are available at https://github.com/kepei1106/DecompEval

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9676–9691

(Ke et al., 2022; Zhong et al., 2022).

To deal with these challenges, we propose a simple yet effective evaluation metric called DecompEval. **Firstly**, to improve the generalization ability, we formulate NLG evaluation as an instruction-style question answering (QA) task, and utilize instruction-tuned pre-trained language models (Chung et al., 2022) to solve this task without training on task-specific data. The instruction-style question consists of an instruction, the input of NLG evaluation, and a yes/no question, e.g., "*Answer the following yes/no question ... Is this a coherent response given the dialogue history?*" for the evaluation of coherence in dialogue generation, where the specific evaluation input is omitted. **Secondly**, we propose a question decomposition strategy to make the evaluation process more interpretable, instead of directly making instruction-tuned PLMs answer the original question. This strategy decomposes the question into the subquestions which sequentially evaluate the corresponding dimension of each sentence in the generated texts. Then, we recompose these subquestions with their answers generated by the PLM as evidence to make the PLM answer the original question, which is used to compute the final evaluation result. The evidence can promote the understanding of the evaluation process by indicating the potential problematic sentences that affect the evaluation score.

Our main contributions are as follows:

- We propose an evaluation metric called DecompEval, which formulates NLG evaluation as an instruction-style QA task, and solves it with instruction-tuned PLMs via question decomposition.

- We conduct experiments on the benchmark datasets for evaluating text summarization and dialogue generation. Experimental results show that DecompEval can achieve state-of-the-art performance in untrained metrics.

- We empirically show that DecompEval can generalize to other evaluation dimensions and tasks (such as data-to-text generation) better than all the baselines, while improving the interpretability via decomposed subquestions with their answers.

2 Related Work

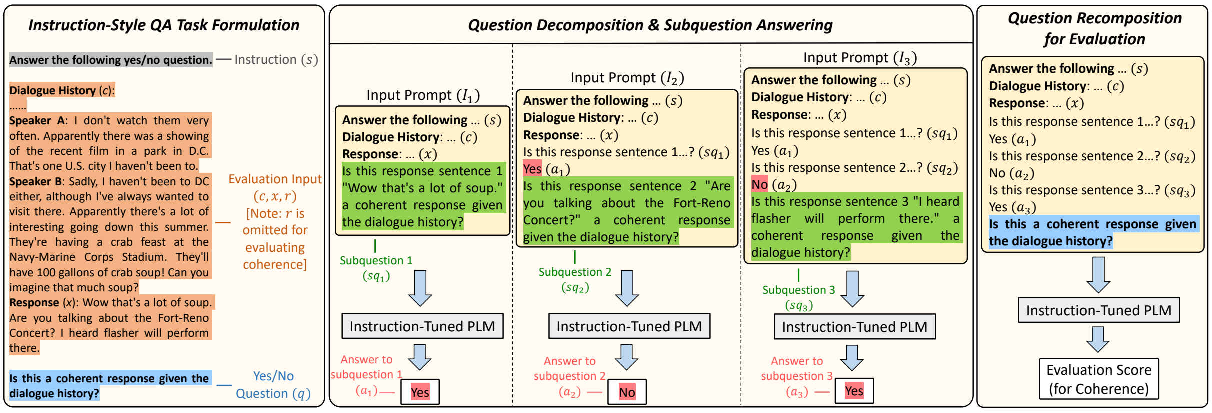
2.1 Evaluation for Language Generation

Evaluation is a long-standing task in the field of NLG (Celikyilmaz et al., 2020), which becomes more critical with the rapid development of PLMs. There are two main categories of automatic evaluation metrics, i.e., untrained and trained metrics (Sai et al., 2020). Untrained metrics without training on specific datasets of evaluation tasks or related tasks aim to measure the relationship among source texts, generated texts, and reference texts via n-gram overlap (Papineni et al., 2002; Banerjee and Lavie, 2005; Lin, 2004), semantic similarity (Zhao et al., 2019; Zhang et al., 2020b), or language modeling / masked language modeling scores (Yuan et al., 2021; Ke et al., 2022; Colombo et al., 2022). In comparison, trained metrics are commonly trained on the evaluation datasets to fit human scores (Shen et al., 2017; Sellam et al., 2020) or distinguish human-written texts from negative samples (Guan and Huang, 2020; Zhong et al., 2022), aiming to achieve higher correlations with human judgments on specific datasets. Among these metrics, there are some similar works which re-frame NLG evaluation as QA tasks and adopt the generated answers or generation probabilities as evaluation results (Deutsch et al., 2021; Zhong et al., 2022).

The most similar work to our method is UniEval (Zhong et al., 2022). UniEval re-frames NLG evaluation as a Boolean QA task and trains the evaluation model on the pseudo data constructed from the evaluation dataset and other related datasets in a unified Boolean QA format. Compared with UniEval, our method is untrained since we transform NLG evaluation to an instruction-style QA task that can be solved by instruction-tuned PLMs without further training. Also, our method can provide some evidence (i.e., the answers to decomposed subquestions) to interpret how the model reaches the evaluation result, instead of only providing a final evaluation score.

2.2 Instruction-Tuned Pre-Trained Models

Instruction learning (Weller et al., 2020) which trains PLMs to follow human instructions has attracted much attention recently since it shows the strong zero-shot cross-task generalization ability. To improve instruction understanding, existing works adopt instruction tuning (Wei et al., 2022) which trains PLMs on massive tasks described


Instruction-Style QA Task Formulation
Instruction (s)
Evaluation Input (c, x, r) [Note: r is omitted for evaluating coherence]
Yes/No Question (q)
Question Decomposition & Subquestion Answering
Input Prompt (I_1)
Input Prompt (I_2)
Input Prompt (I_3)
Subquestion 1 (sq_1)
Subquestion 2 (sq_2)
Subquestion 3 (sq_3)
Instruction-Tuned PLM
Answer to subquestion 1 (a_1)
Yes
Answer to subquestion 2 (a_2)
No
Answer to subquestion 3 (a_3)
Yes
Question Recomposition for Evaluation
Instruction-Tuned PLM
Evaluation Score (for Coherence)

Answer the following yes/no question.

Dialogue History (c):
……
Speaker A: I don't watch them very often. Apparently there was a showing of the recent film in a park in D.C. That's one U.S. city I haven't been to.
Speaker B: Sadly, I haven't been to DC either, although I've always wanted to visit there. Apparently there's a lot of interesting going down this summer. They're having a crab feast at the Navy-Marine Corps Stadium. They'll have 100 gallons of crab soup! Can you imagine that much soup?
Response (x): Wow that's a lot of soup. Are you talking about the Fort-Reno Concert? I heard flasher will perform there.

Is this a coherent response given the dialogue history?

Answer the following … (s)
Dialogue History: … (c)
Response: … (x)
Is this response sentence 1 "Wow that's a lot of soup." a coherent response given the dialogue history?

Answer the following … (s)
Dialogue History: … (c)
Response: … (x)
Is this response sentence 1…? (sq_1)
Yes (a_1)
Is this response sentence 2 "Are you talking about the Fort-Reno Concert?" a coherent response given the dialogue history?

Answer the following … (s)
Dialogue History: … (c)
Response: … (x)
Is this response sentence 1…? (sq_1)
Yes (a_1)
Is this response sentence 2…? (sq_2)
No (a_2)
Is this response sentence 3 "I heard flasher will perform there." a coherent response given the dialogue history?

Answer the following … (s)
Dialogue History: … (c)
Response: … (x)
Is this response sentence 1…? (sq_1)
Yes (a_1)
Is this response sentence 2…? (sq_2)
No (a_2)
Is this response sentence 3…? (sq_3)
Yes (a_3)
Is this a coherent response given the dialogue history?

Figure 1: The overview of DecompEval. We take the evaluation of coherence in dialogue generation as an example. **Left**: The input of evaluation is formulated as an instruction-style question, which contains an instruction, a tuple of evaluation inputs, and a yes/no question about the quality of generated responses. **Medium**: The instruction-style question is decomposed into subquestions according to sentences. At each step, the instruction-tuned PLM generates an answer to the current subquestion based on the input prompt. Then, the answer becomes the constituent of the input prompt at the next step. **Right**: The instruction-tuned PLM recomposes all the subquestions with their answers to answer the original question and acquire the evaluation result.

via instructions with multi-task learning, such as FLAN (Wei et al., 2022; Chung et al., 2022), T0 (Sanh et al., 2022), and InstructGPT (Ouyang et al., 2022). Other works systematically study instruction tuning in specific areas such as dialogue systems (Gupta et al., 2022) and multi-modal learning (Xu et al., 2022).

In comparison, our work is the first to explore the potential of instruction-tuned PLMs in the evaluation of NLG without further training. We show that equipped with well-designed input prompts and suitable question decomposition, instruction-tuned PLMs can sequentially measure the quality of each sentence and finally recompose all the subquestions with their answers to obtain surprisingly great evaluation results in an unsupervised fashion.

3 Method

3.1 Task Definition and Model Overview

Given the context c, the model-generated text x, and the reference text r, our goal is to acquire the evaluation results from different individual dimensions, respectively. The context contains different contents in various NLG tasks. Also, the context and the reference may be omitted, which depend on the evaluation task and dimension. We assume that the generated text consists of n sentences, i.e., $x = (x_1, x_2, \cdots, x_n)$.

As shown in Figure 1, our main idea is to formulate NLG evaluation as an instruction-style QA task and solve this task with instruction-tuned PLMs via question decomposition. Our proposed method consists of three steps. First of all, we transform the input of NLG evaluation into an instruction-style question which contains an instruction s, the input of evaluation tasks (c, x, r), and a yes/no question q for each dimension (§3.2). Then, we decompose this question into the subquestions $\{sq_t\}_{t=1}^{n}$, which evaluate each sentence $x_t (1 \leq t \leq n)$ in the generated text x respectively and acquire the answers $\{a_t\}_{t=1}^{n}$ to these subquestions via the instruction-tuned PLM P_θ (§3.3). The answer to each subquestion is appended to the input prompt of the PLM, which may help to solve subsequent subquestions as in-context examples. Finally, we recompose all the subquestions with their answers as evidence and make the instruction-tuned PLM answer the original question, which can be used to compute the evaluation result (§3.4).

3.2 Instruction-Style QA Task Formulation

To improve the generalization ability of evaluation metrics, we formulate NLG evaluation as an instruction-style QA task that can be solved by instruction-tuned PLMs in an unsupervised fashion. As shown in Figure 1, the instruction-style question contains three parts:

- **Instruction**: The design of instructions depends on the data format of instruction-tuned PLMs. In this paper, we adopt yes/no questions (Zhong et al., 2022) to measure the quality of generated texts. Thus, we follow

Chung et al. (2022) to devise the instruction as $s =$"*Answer the following yes/no question.*".

- **Evaluation Input**: The original input (c, x, r) for NLG evaluation mentioned in §3.1 are incorporated with task-specific descriptive texts. For example, we add the text "*dialogue history:*", "*response:*", and "*reference:*" before c, x, and r respectively for evaluating dialogue generation.

- **Yes/No Question**: We finally devise a yes/no question to assess the specific dimension of generated texts. For example, the yes/no question assessing the coherence of generated texts in dialogue generation is $q =$"*Is this a coherent response given the dialogue history?*".

3.3 Question Decomposition and Subquestion Answering

To interpret how the model predicts the evaluation score, we devise a question decomposition strategy inspired by the existing works in the QA community (Min et al., 2019; Perez et al., 2020; Zhou et al., 2023), rather than force the instruction-tuned PLM to answer the original question directly. This strategy splits the generated text based on sentences and sequentially queries the quality of each sentence via subquestions. The subquestions with their answers generated by the PLM are expected to act as evidence to illustrate how the PLM arrives at the final evaluation score. We simply select sentences as the decomposition criterion instead of using external off-the-shelf models (Perez et al., 2020; Deutsch et al., 2021) because sentences are shown to be important basic units for deriving the evaluation result of the whole generated text (Amplayo et al., 2023).

Specifically, to answer the subquestion $sq_t (1 \leq t \leq n)$ for measuring the quality of the t-th sentence x_t, we combine the instruction s, the evaluation input (c, x, r), the previous subquestions with their answers $\{(sq_j, a_j)\}_{j=1}^{t-1}$, and the current subquestion sq_t as the input prompt $I_t = \left(s, c, x, r, \{(sq_j, a_j)\}_{j=1}^{t-1}, sq_t\right)$. Then, we compare the generation probability of "*yes*" / "*no*" from the instruction-tuned PLM to determine the answer:

$$\begin{aligned} a_t &= \begin{cases} \text{yes}, & P_\theta(\text{yes}|I_t) > P_\theta(\text{no}|I_t) \\ \text{no}, & P_\theta(\text{yes}|I_t) \leq P_\theta(\text{no}|I_t) \end{cases} \\ t &= 1, 2, \cdots, n \end{aligned} \quad (1)$$

The answer a_t is appended to the current input prompt I_t, which becomes the in-context examples of I_{t+1} helping to solve the next subquestion sq_{t+1}. All these subquestions with their answers can serve as evidence to improve the interpretability by indicating potential low-quality sentences in the generated text that affect the evaluation score.

3.4 Question Recomposition for Evaluation

To recompose all the subquestions with their answers to acquire the final evaluation result, we append the original yes/no question mentioned in §3.2 to the end of the last subquestion and its answer. The instruction-tuned PLM is expected to leverage all these information as evidence to answer the original question and obtain the evaluation result.

Specifically, given the instruction s, the evaluation input (c, x, r), all the subquestions with their answers $\{(sq_t, a_t)\}_{t=1}^{n}$, and the original question q as the input prompt, we compute the evaluation score using the generation probability of answer words (i.e., yes and no) from the instruction-tuned PLM (Ke et al., 2022; Zhong et al., 2022):

$$f(l) = P_\theta(l|s, c, x, r, \{(sq_t, a_t)\}_{t=1}^{n}, q) \quad (2)$$

$$score = \frac{f(l = \text{yes})}{f(l = \text{yes}) + f(l = \text{no})} \quad (3)$$

4 Experiment

4.1 Dataset

We follow Zhong et al. (2022) to adopt two benchmark datasets to test the performance of DecompEval. The statistics of these datasets are shown in Table 1.

SummEval (Fabbri et al., 2021): This dataset is a benchmark for evaluation metrics of text summarization. It covers the generated summaries from recent summarization models on the CNN/DailyMail (CNNDM) dataset (Hermann et al., 2015). For each generated summary, it provides the human scores from four dimensions including fluency, coherence, consistency, and relevance.

Topical-Chat (Gopalakrishnan et al., 2019): This dataset is a benchmark for knowledge-grounded dialogue generation. Mehri and Eskénazi (2020) collects human annotations for the models trained on Topical-Chat. For each generated response, it provides the human scores from five dimensions[2]

[2] We use the description of dimensions in the existing work (Zhong et al., 2022) for fair comparison, which is slightly different from the original paper (Mehri and Eskénazi, 2020).

Dataset	Task	#Samples	#Dimensions	Length
SummEval	Text Summarization	1,600	4	63.7
Topical-Chat	Dialogue Generation	360	5	22.9

Table 1: Statistics of the benchmark datasets, including the task, the number of samples / dimensions, and the average length of generated texts.

including naturalness, coherence, engagingness, groundedness, and understandability. Following Zhong et al. (2022), we use the first four dimensions in the main result (§4.4) and the last dimension to test the generalization ability (§4.5).

4.2 Implementation Detail

We choose FLAN-T5 (Chung et al., 2022) as our base model, which is obtained by training T5 (Raffel et al., 2020) on 1.8K tasks described via instructions[3]. We use FLAN-T5-XL with 3B parameters in the main result and also explore other model scales in §4.8. We follow Zhong et al. (2022) to set the input length to be 1,024. We design the input prompts based on the data formats of FLAN-T5, the evaluation tasks and dimensions. More details about the specific design of input prompts for each dataset / dimension and the sensitivity analysis are included in Appendix A.

As for the evaluation on two datasets, we directly compute summary-level / turn-level evaluation scores for SummEval / Topical-Chat based on our method in most of the dimensions, respectively, except fluency / consistency on SummEval and engagingness on Topical-Chat. For these dimensions, we follow Zhong et al. (2022) to obtain the evaluation scores via averaging (for fluency / consistency on SummEval) (Laban et al., 2022) or cumulating (for engagingness on Topical-Chat) (Deng et al., 2021) individual evaluation results of constituent sentences for fair comparison.

4.3 Baseline

We choose several state-of-the-art untrained and trained metrics as our baselines:

MoverScore (Zhao et al., 2019): This metric relies on Earth Mover's Distance (Rubner et al., 2000) between generated texts and reference texts based on the contextual representations from PLMs.

BERTScore (Zhang et al., 2020b): This metric computes the similarity between generated texts and reference texts based the contextual representations from BERT (Devlin et al., 2019).

USR (Mehri and Eskénazi, 2020): This metric combines the evaluation results of masked language models and dialogue retrieval models which are trained on the dialogue evaluation dataset.

BARTScore (Yuan et al., 2021): This metric utilizes the generation probabilities of BART (Lewis et al., 2020) to measure the relationship among source texts, generated texts, and reference texts with different inputs and outputs. We use two variants **BARTScore** and **BARTScore (CNNDM)** in the original paper. The latter adopts BART fine-tuned on the CNNDM dataset as the base model.

CTRLEval (Ke et al., 2022): This metric formulates evaluation dimensions as multiple text infilling tasks and uses the ensemble of generation probabilities from PEGASUS (Zhang et al., 2020a) as the evaluation results.

UniEval (Zhong et al., 2022): This metric reframes NLG evaluation as a Boolean QA task. It conducts multi-task learning on the related datasets and continual learning on the dimensions of the evaluation dataset with a unified QA format. We use two variants **UniEval (Summ)** and **UniEval (Dial)** in the original paper, which are trained on all the dimensions of SummEval and the first four dimensions of Topical-Chat, respectively.

In addition, we also select traditional evaluation metrics based on n-gram overlap like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004) as baselines. We directly re-print the experimental results of baselines if their original papers adopt the same benchmark datasets as ours. Otherwise, we implement the baselines based on the codes and model parameters released by the original papers.

4.4 Main Result

Following Liu et al. (2021) and Zhong et al. (2022), we adopt summary-level Spearman (ρ) and Kendall (τ) correlation coefficients between human judgments and automatic metrics to assess the performance on the SummEval dataset. The results in Table 2 show that DecompEval achieves state-of-the-art performance in untrained metrics, indicating the effectiveness of our proposed instruction-style QA formulation and question decomposition method. Especially, DecompEval can even beat the best-performing trained metric UniEval (Summ) in the dimension of consistency, which shows the po-

[3] Although the instruction-tuning datasets of FLAN-T5 cover the CNNDM dataset (Chung et al., 2022), they do not include the generated summaries with human evaluation scores, ensuring no data leak in the experiment.

Dimension	Coherence		Consistency		Fluency		Relevance	
Metric	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Trained Metric (w/ Training on Data of Evaluation Tasks or Related Tasks)								
BARTScore (CNNDM)	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273
UniEval (Summ)	0.575	0.442	0.446	0.371	0.449	0.371	0.426	0.325
Untrained Metric (w/o Training on Data of Evaluation Tasks or Related Tasks)								
ROUGE-1	0.167	0.126	0.160	0.130	0.115	0.094	0.326	0.252
ROUGE-2	0.184	0.139	0.187	0.155	0.159	0.128	0.290	0.219
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237
MoverScore	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244
BERTScore	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243
BARTScore	0.322	0.250	0.311	0.256	0.248	0.203	0.264	0.197
CTRLEval	0.217	0.164	0.301	0.247	0.132	0.107	0.196	0.152
DecompEval (Ours)	**0.341**	**0.256**	**0.455**	**0.378**	**0.285**	**0.233**	**0.355**	**0.276**

Table 2: Summary-level Spearman (ρ) and Kendall (τ) correlations of coherence, consistency, fluency, and relevance on the SummEval dataset. The highest correlation for each dimension achieved by untrained metrics is **bold**, while the highest correlation overall is underlined.

Dimension	Naturalness		Coherence		Engagingness		Groundedness	
Metric	r	ρ	r	ρ	r	ρ	r	ρ
Trained Metric (w/ Training on Data of Evaluation Tasks or Related Tasks)								
USR	0.337	0.325	0.416	0.377	0.456	0.465	0.222	0.447
UniEval (Dial)	0.444	0.514	0.595	0.613	0.557	0.605	0.536	0.575
Untrained Metric (w/o Training on Data of Evaluation Tasks or Related Tasks)								
BLEU-1	0.161	0.133	0.210	0.223	0.314	0.334	0.289	0.303
BLEU-4	0.180	0.175	0.131	0.235	0.232	0.316	0.213	0.310
ROUGE-L	0.176	0.146	0.193	0.203	0.295	0.300	0.310	0.327
METEOR	0.212	0.191	0.250	0.302	0.367	0.439	0.333	0.391
MoverScore	0.169	0.170	0.247	0.259	0.275	0.269	0.198	0.147
BERTScore	0.226	0.209	0.214	0.233	0.317	0.335	0.291	0.317
BARTScore	0.287	0.266	0.251	0.225	0.411	0.406	0.226	0.205
CTRLEval	0.303	0.254	0.337	0.313	0.422	0.412	0.242	0.251
DecompEval (Ours)	**0.410**	**0.435**	**0.434**	**0.435**	**0.453**	**0.467**	**0.646**	**0.659**

Table 3: Turn-level Pearson (r) and Spearman (ρ) correlations of naturalness, coherence, engagingness, and groundedness on the Topical-Chat dataset. The highest correlation for each dimension achieved by untrained metrics is **bold**, while the highest correlation overall is underlined.

tential of instruction-tuned PLMs in the evaluation of generated texts.

We also conduct experiments on the Topical-Chat dataset and report turn-level Pearson (r) / Spearman (ρ) correlation coefficients in Table 3 as the existing works (Mehri and Eskénazi, 2020; Zhong et al., 2022) do. Similarly, DecompEval beats all the untrained baselines and even outperforms the trained baseline USR in most of the dimensions. This indicates that DecompEval can successfully adapt to the evaluation of dialogue generation without training on specific datasets. We also find that DecompEval can outperform UniEval (Dial) in the dimension of groundedness. We conjecture that DecompEval may be good at measuring the consistency between generated texts and contexts, thereby performing extremely well on consistency in text summarization and groundedness in dialogue generation.

4.5 Generalization Ability

Generalization ability is essential because new evaluation dimensions and tasks may emerge without sufficient data. Thus, we study whether DecompEval can generalize at the dimension / task level better than untrained and trained baselines.

4.5.1 Generalization to Other Dimensions

To compare the performance of DecompEval and untrained / trained baselines on other dimensions, we follow Zhong et al. (2022) to adopt the dimension of understandability on the Topical-Chat

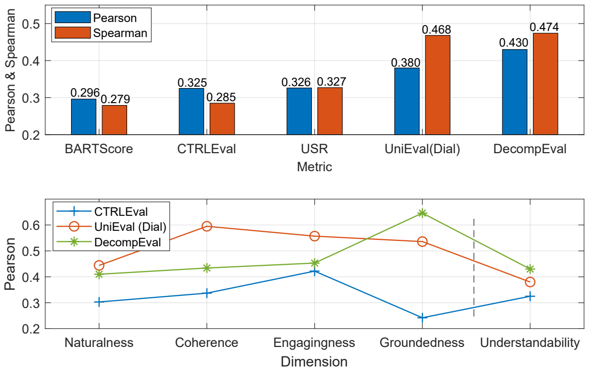


Figure 2: **Top**: Pearson and Spearman correlations of different metrics in the dimention of understandability. **Bottom**: Pearson correlation of CTRLEval, UniEval (Dial), and DecompEval in all the five dimensions of Topical-Chat.

dataset to conduct experiments.

The results in the top of Figure 2 show that DecompEval can outperform all the competitive untrained / trained baselines and achieve best performance in the dimension of understandability, which shows its strong dimension-level generalization ability. From the bottom of Figure 2, we can observe that DecompEval maintains stable performance in all these dimensions. In comparison, the trained baseline UniEval (Dial) which is trained on the first four dimensions of Topical-Chat except understandability cannot surpass DecompEval in the evaluation of understandability. The performance of UniEval (Dial) also degrades obviously in understandability compared with the other dimensions, which demonstrates the potential side-effect of over-fitting to specific dimensions.

4.5.2 Generalization to Other NLG tasks

To investigate how DecompEval performs compared with untrained / trained baselines in other NLG tasks in addition to text summarization and dialogue generation, we follow Yuan et al. (2021) and Zhong et al. (2022) to adopt two data-to-text generation datasets SFRES and SFHOT (Wen et al., 2015). These two datasets cover generated texts from structured data in the domain of restaurants and hotels. For each generated text, they provide human scores from two dimensions, i.e., naturalness and informativeness. The number of samples in SFRES / SFHOT is 1,181 / 875, respectively.

The results are shown in Table 4. Our proposed metric DecompEval can still achieve state-of-the-art performance in untrained metrics and outperform the trained baselines in most of the dimensions. Thus, we believe that DecompEval can successfully improve the generalization ability to multiple NLG tasks via the full utilization of the instruction-tuned PLM without further training. We also illustrate the average of Spearman correlation coefficients in all the dimensions of each dataset in Table 5. Compared with our proposed metric, UniEval (Summ) and UniEval (Dial), as the best-performing trained metrics on the SummEval and Topical-Chat datasets, respectively, obtain obviously worse performance on the evaluation datasets which they are not trained on, indicating limited task-level generalization ability.

Dataset	SFRES		SFHOT	
Metric	Nat.	Info.	Nat.	Info.
Trained Metric				
BARTScore (CNNDM)	0.289	0.238	0.288	0.235
UniEval (Summ)	0.333	0.225	0.320	0.249
UniEval (Dial)	0.291	0.194	0.291	0.196
Untrained Metric				
ROUGE-1	0.170	0.115	0.196	0.118
ROUGE-L	0.169	0.103	0.186	0.110
MoverScore	0.190	0.153	0.242	0.172
BERTScore	0.219	0.156	0.178	0.135
BARTScore	0.200	0.164	0.165	0.158
CTRLEval	0.195	0.177	0.121	0.158
DecompEval (Ours)	**0.345**	**0.242**	**0.316**	**0.302**

Table 4: Spearman correlation of naturalness (Nat.) and informativeness (Info.) on data-to-text generation datasets. The highest correlation for each dimension achieved by untrained metrics is **bold**, while the highest correlation overall is underlined.

Dataset	Ours vs. UniEval (Summ)	Ours vs. UniEval (Dial)
SummEval	🟩 0.359 vs. 0.474	🟥 0.359 vs. 0.305
Topical-Chat	🟥 0.499 vs. 0.315	🟩 0.499 vs. 0.577
SFRES	🟥 0.293 vs. 0.279	🟥 0.293 vs. 0.243
SFHOT	🟥 0.309 vs. 0.285	🟥 0.309 vs. 0.244

Table 5: Comparison of Spearman correlation averaged over all the dimensions in each dataset. Red indicates that our metric is better while green means the opposite.

4.6 Interpretability

To verify whether the subquestions with their answers are reliable evidence to interpret the evaluation score, we conduct human evaluation on the generated answers to subquestions. We randomly select 200 subquestions from each dimension of the Topical-Chat dataset. Three annotators are hired to answer these subquestions with yes or no according to the evaluation input, where the human-annotated labels are determined via majority voting. The

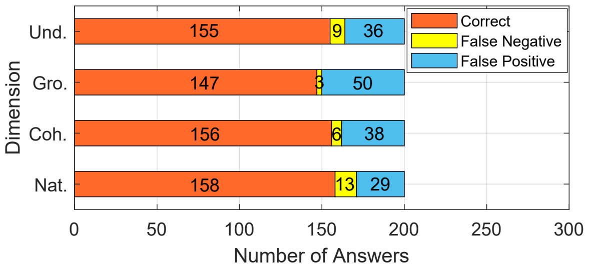


Figure 3: Human evaluation on subquestion answering in naturalness (Nat.), coherence (Coh.), groundedness (Gro.), and understandability (Und.) of Topical-Chat.

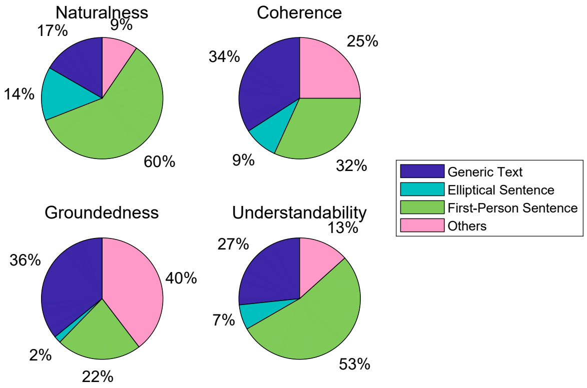


Figure 4: Type distributions of generated texts in the error cases of human evaluation.

human-annotated labels are used as ground-truth labels to measure the quality of generated answers.

The results in Figure 3 show that the accuracy of each dimension is above 0.7, indicating reasonable performance of subquestion answering which serves as interpretable evidence. We manually check the error cases and find that they include three typical types of generated texts, i.e., generic texts (Li et al., 2016), elliptical sentences, and first-person sentences. The type distributions of generated texts in the error cases are shown in Figure 4. We can observe that generic texts (such as "*That is true.*") dominate the generated texts in the error cases of coherence / groundedness, while first-person sentences (such as "*I think ...*") appear more frequently in those of naturalness / understandability. These two types of generated texts are mostly not contradictory to the evaluation input, thereby being commonly recognized by our metric. However, generic texts can only provide limited information while first-person sentences may contain irrelevant contents regarding the evaluation input. Thus, annotators tend to regard them as low-quality ones.

We also provide the case study in Appendix B to show the interpretable evaluation process of DecompEval.

Metric	Nat.	Coh.	Eng.	Gro.
DecompEval	**0.435**	**0.435**	**0.467**	**0.659**
w/o Instruction	0.427	0.418	0.442	0.651
w/o Decomp. Q&A	0.411	0.399	0.433	0.643
w/ Prefix Yes/No Que.	0.431	0.402	0.461	0.601

Table 6: Spearman correlation of ablation models in naturalness (Nat.), coherence (Coh.), engagingness (Eng.), and groundedness (Gro.) of the Topical-Chat dataset.

4.7 Ablation Study

To further investigate the effectiveness of each part in our metric, we conduct detailed ablation studies. We build the following three ablation models which remove three important parts of input prompts in §3.4, respectively: 1) *w/o Instruction* indicates the model without the instruction s; 2) *w/o Decomp. Q&A* denotes the model without the decomposed subquestions with their answers $\{(sq_t, a_t)\}_{t=1}^{n}$; 3) *w/ Prefix Yes/No Que.* means moving the yes/no question q to the prefix of the evaluation input behind the instruction. We find that our metric without this yes/no question fails to achieve reasonable performance possibly because it contains the information about evaluation tasks and dimensions.

The results are shown in Table 6. We can observe that all these three parts contribute to final performance. The decomposed subquestions with their answers play a more important role in most of the dimensions, indicating their positive impact as evidence on the model performance in addition to the interpretability. As for instructions, the performance of DecompEval without instructions does not degrade obviously. We conjecture that the yes/no question has explicitly conveyed the information to make the instruction-tuned PLM answer with yes or no. Thus, the impact of instructions may be weakened. The position of yes/no questions has also affected the model performance. From the experimental results, the question in the end of input prompts can obtain better performance than that in the middle part.

4.8 Analysis on Model Scale

We further conduct experiments on the scale of base models, which may impact the capacity of following instructions to evaluate generated texts. We choose FLAN-T5-Base and FLAN-T5-Large additionally, and compare their performance with

Base Model	#Param	Nat.	Coh.	Eng.	Gro.
FLAN-T5-Base	250M	0.175	0.206	0.386	0.291
FLAN-T5-Large	780M	0.217	0.165	0.390	0.525
FLAN-T5-XL	3B	**0.435**	**0.435**	**0.467**	**0.659**

Table 7: Spearman correlation of different base models in naturalness (Nat.), coherence (Coh.), engagingness (Eng.), and groundedness (Gro.) of Topical-Chat. #Param means the number of model parameters.

FLAN-T5-XL used in our main experiments.

The results in Table 7 show that the performance of DecompEval improves on most of the dimensions as the number of parameters in the base model increases. We also find that there is a relatively large margin between the performance of FLAN-T5-Base/Large and FLAN-T5-XL, especially in the dimensions of naturalness, coherence, and groundedness. This phenomenon is accordant to the findings of existing works (Chung et al., 2022; Wei et al., 2022), where the zero-shot capacity of instruction following mainly emerges in the models of sufficiently large scales.

5 Discussion

Applicability in Non-English Languages: Although the benchmark datasets in the experiment are mainly in English, our method can be also applied to non-English languages. Since our base model FLAN-T5 has some multilingual ability (Chung et al., 2022), we can design instruction-style questions / subquestions and answer words in the target language to apply DecompEval to non-English evaluation tasks. DecompEval can also adapt to stronger instruction-tuned multilingual PLMs for better applicability in non-English languages. We will further investigate the extensibility of our method to non-English evaluation tasks in the future work.

6 Conclusion

We present an untrained evaluation metric called DecompEval, which formulates NLG evaluation as an instruction-style QA task, and utilizes instruction-tuned PLMs to solve this task via question decomposition. Experimental results show that DecompEval achieves state-of-the-art performance in untrained metrics, which also exhibits better dimension-level / task-level generalization ability than trained metrics and improves the interpretability.

Limitations

The limitation of our work includes the following aspects:

1) The instruction-style question which measures the quality of generated texts from different dimensions still needs manual design. Although the questions in our experiment have already involved typical dimensions in text summarization, dialogue generation, and data-to-text generation, we admit that it is hard to cover all the dimensions in various NLG tasks. We believe that this is not a severe problem because we can refer to the definition and human annotation instructions (Mehri and Eskénazi, 2020) of each dimension, which are commonly formulated as questions. We leave the exploration of automatically constructing instruction-style questions for multiple dimensions of NLG evaluation as future work.

2) Due to the limitation of computational resources, the largest base model used in our experiment is FLAN-T5-XL with 3B parameters. Since the ability of instruction following is related to the model scale (Wei et al., 2022), we leave the exploration of adopting larger instruction-tuned PLMs such as FLAN-T5-XXL and OPT-IML (Iyer et al., 2022) as future work.

Acknowledgements

This work was supported by the NSFC project (Key project with No. 61936010). This work was also supported by the Guoqiang Institute of Tsinghua University, with Grant No. 2020GQG0005.

References

Reinald Kim Amplayo, Peter J Liu, Yao Zhao, and Shashi Narayan. 2023. Smart: Sentences as basic units for text evaluation. In *The Eleventh International Conference on Learning Representations*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Pierre Jean A Colombo, Chloé Clavel, and Pablo Piantanida. 2022. Infolm: A new metric to evaluate summarization & data2text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10554–10562.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Trans. Assoc. Comput. Linguistics*, 9:774–789.

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*, pages 4171–4186.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *20th Annual Conference of the International Speech Communication Association*, pages 1891–1895.

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*, pages 9157–9166.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P Bigham. 2022. Instructdial: Improving zero and few-shot generalization in dialogue through instruction tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. CTRLEval: An unsupervised reference-free metric for evaluating controlled text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Trans. Assoc. Comput. Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

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

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. Explainaboard: An explainable leaderboard for NLP. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 280–289.

Shikib Mehri and Maxine Eskénazi. 2020. 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.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of* *the Association for Computational Linguistics*, pages 6097–6109.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

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

Ethan Perez, Patrick S. H. Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8864–8880.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.*, 40(2):99–121.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Trans. Assoc. Comput. Linguistics*, 8:810–827.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6830–6841.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations*.

Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. 2020. Learning from task descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Peihao Su, David Vandyke, and Steve J. Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2022. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11328–11339.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 563–578.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Prompt Design

We show the specific prompt design for each dimension of SummEval, Topical-Chat, and SFRES / SFHOT in Table 8, 9, and 10, respectively. The instruction used in all the datasets is $s =$"*Answer the following yes/no question.*", as mentioned in §3.2. We refer to the definition and human annotation instructions of each dimension (Fabbri et al., 2021; Mehri and Eskénazi, 2020; Wen et al., 2015) as well as the existing works on QA for evaluation (Deutsch et al., 2021; Zhong et al., 2022) to design evaluation inputs and yes/no questions. The format of subquestions is similar to yes/no questions, where the sentence to be measured is added to the middle part.

To investigate the sensitivity of input prompts, we construct seven grammatical yes/no questions for each dimension of Topical-Chat, covering the original one and three types of lexical variations, i.e., auxiliary verb replacement, synonym replacement, and word reordering. For example, the original yes/no question for naturalness in Table 9 is "*Is this response natural to the dialogue history?*". After auxiliary verb replacement, the question may start with another auxiliary verb, such as "*Does this response have a natural body to the dialogue history?*". Similarly, after synonym replacement, the question may have some words which are replaced with their synonyms, such as "*Is this response natural given the dialogue history?*". As for word reordering, the question may be composed of reordered words, such as "*Is this a natural response to the dialogue history?*". Note that the subquestions are perturbed accordingly. Then, we illustrate the mean value and standard deviation over the original prompt and perturbed prompts of each dimension in Figure 5, showing the stable performance of DecompEval faced with variations.

B Case Study

We provide evaluation cases on the Topical-Chat and SummEval datasets in Table 11 and 12, respectively. We can observe that DecompEval can provide the evaluation scores which are the most accordant to human scores. Also, the subquestions with their answers can act as evidence to indicate the potential low-quality sentence which impacts the overall quality. For example, in Table 11, the second sentence which mentions the concert seems not to be coherent to the topic in the dialogue history (i.e., the crab feast at a stadium). Similarly, in Table 12, the third sentence about patient satisfaction is not relevant to the reference. In comparison, the evaluation scores of other metrics deviate from human scores, while they cannot provide evidence to demonstrate how they predict the evaluation scores.

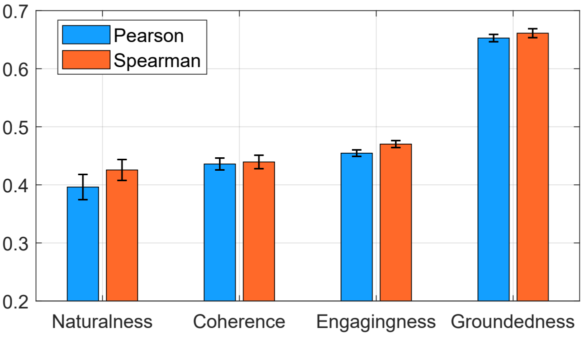


Figure 5: Mean values and standard deviations of Pearson and Spearman correlations over the original prompt and perturbed prompts on the Topical-Chat dataset.

C Analysis on Decomposition Strategy

To judge how useful our decomposed subquestions with generated answers are for interpreting final evaluation scores, we ask the same three annotators to assign an interpretability score to selected samples in §4.6. We adopt a 1-3 Likert scale, where 1 / 2 / 3 means that the decomposition can hardly / partly / comprehensively help understand how the model reaches final scores, respectively. The average interpretability scores over all the selected samples are 2.84 / 2.76 / 2.74 / 2.67 for naturalness / coherence / groundedness / understandability, respectively, showing that our decomposition strategy based on sentences is mostly useful for interpreting final evaluation scores of multiple dimensions.

D Experimental Detail

D.1 License of Datasets and Models

The licenses of datasets and base models used in our experiments include MIT for the SummEval dataset and Apache-2.0 for the Topical-Chat dataset and the FLAN-T5 model.

D.2 Implementation Detail

We use NLTK[4] to split generated texts into sentences for the construction of subquestions. As for the computation of Pearson, Spearman, and Kendall correlation coefficients, we use the APIs from SciPy[5].

[4] https://www.nltk.org

[5] https://scipy.org

Dimension	Evaluation Input	Yes/No Question	Subquestion
Coherence	document: c summary: x	Is this a coherent summary to the document?	Is this summary sentence t x_t a coherent summary to the document?
Consistency	claim: x document: c	Is this claim consistent with the document?	Is this claim sentence t x_t consistent with the document?
Fluency	paragraph: x	Is this a fluent paragraph?	Is this paragraph sentence t x_t a fluent paragraph?
Relevance	summary: x reference: r	Is this summary relevant to the reference?	Is this summary sentence t x_t relevant to the reference?

Table 8: Input prompt design for each dimension of the SummEval dataset, including the evaluation inputs (c, x, r), yes/no questions (q), and decomposed subquestions $(\{sq_t\}_{t=1}^n)$.

Dimension	Evaluation Input	Yes/No Question	Subquestion
Naturalness	dialogue history: c_{his} response: x	Is this response natural to the dialogue history?	Is this response sentence t x_t natural to the dialogue history?
Coherence	dialogue history: c_{his} response: x	Is this a coherent response given the dialogue history?	Is this response sentence t x_t a coherent response given the dialogue history?
Engagingness	dialogue history: c_{his} fact: c_{fact} response: x	Is this an engaging response according to the dialogue history and fact?	Is this response sentence t x_t an engaging response according to the dialogue history and fact?
Groundedness	response: x fact: c_{fact}	Is this response consistent with knowledge in the fact?	Is this response sentence t x_t consistent with knowledge in the fact?
Understandability	dialogue history: c_{his} response: x	Is this an understandable response given the dialogue history?	Is this response sentence t x_t an understandable response given the dialogue history?

Table 9: Input prompt design for each dimension of the Topical-Chat dataset, including the evaluation inputs (c, x, r), yes/no questions (q), and decomposed subquestions $(\{sq_t\}_{t=1}^n)$. Note that Topical-Chat is a knowledge-grounded dialogue generation dataset, where the context c contains dialogue histories c_{his} and knowledge facts c_{fact}.

Dimension	Evaluation Input	Yes/No Question	Subquestion
Naturalness	utterance: x	Is this a fluent utterance?	Is this utterance sentence t x_t a fluent utterance?
Informativeness	sentence: x reference: r	Is this sentence informative according to the reference?	Is this sentence t x_t informative according to the reference?

Table 10: Input prompt design for each dimension of the SFRES / SFHOT dataset, including the evaluation inputs (c, x, r), yes/no questions (q), and decomposed subquestions $(\{sq_t\}_{t=1}^n)$.

D.3 Inference Time

The inference time on the SummEval / Topical-Chat dataset is about 28 / 5 minutes, respectively. We test our model on 1 NVIDIA A100 GPU.

D.4 Human Evaluation

The annotation instruction of human evaluation in §4.6 contains two main parts: 1) A subquestion with its corresponding instruction and evaluation input in the same format as Figure 1; 2) An explanation of NLG tasks and dimensions to be measured, which is re-printed from the original paper about benchmark datasets (Mehri and Eskénazi, 2020). In addition, all the other contents shown to annotators are from the original dataset of Topical-Chat (Gopalakrishnan et al., 2019). We manually check these contents before annotation to avoid potential risks.

We recruit three graduate students as annotators to complete this task. We pay each annotator $0.07 for every subquestion. The payment is determined based on the difficulty of tasks and the length of subquestions.

Dialogue History	... Speaker A: I don't watch them very often. Apparently there was a showing of the recent film in a park in D.C. That's one U.S. city I haven't been to. Speaker B: Sadly, I haven't been to DC either, although I've always wanted to visit there. Apparently there's a lot of interesting going down this summer. They're having a crab feast at the Navy-Marine Corps Stadium. They'll have 100 gallons of crab soup! Can you imagine that much soup?
Generated Response	Wow that's a lot of soup. Are you talking about the Fort-Reno Concert? I heard flasher will perform there.
Evaluation Dimension	Coherence
Human Score (1-3)	2.667
MoverScore (0-1)	0.506
BARTScore (<0)	-3.867
CTRLEval (<0)	-4.768
UniEval (Dial) (0-1)	0.999
DecompEval (0-1) w/ Evidence	0.855 Is this response sentence 1 "Wow that's a lot of soup." a coherent response given the dialogue history? **Yes** Is this response sentence 2 "Are you talking about the Fort-Reno Concert?" a coherent response given the dialogue history? **No** Is this response sentence 3 "I heard flasher will perform there." a coherent response given the dialogue history? **Yes**

Table 11: Case study on the evaluation of coherence in the Topical-Chat dataset. The content in the bracket indicates the scale of evaluation scores in each metric, where higher scores mean better quality. The evidence of DecompEval denotes the subquestions with their answers.

Document	A southern Iowa chiropractor accused of accepting sex as payment for his services and performing exorcisms on patients has surrendered his state license ...
Generated Summary	A chiropractor in iowa has surrendered his license to practice and admitted to swapping services for sex and performing exorcisms on some patients. Manuel also recommended that patients stop taking medication no longer exist before he can resume practicing chiropractic in the state. The disgraced chiropractor received a perfect five out of five stars in patient satisfaction.
Reference Summary	Charles Manuel of Lamoni, Iowa admitted to a review board that he traded sexual favors for his services. Manuel also fessed up to performing exorcisms and to telling patients to stop taking medications prescribed to them by a medical doctor. The Iowa Board of Chiropractic required Manuel to pledge he would not apply for reinstatement of the license, but only for 10 years.
Evaluation Dimension	Relevance
Human Score (1-5)	3.667
MoverScore (0-1)	0.546
BARTScore (<0)	-5.188
CTRLEval (<0)	-2.912
UniEval (Summ) (0-1)	0.060
DecompEval (0-1) w/ Evidence	0.586 Is this summary sentence 1 "A chiropractor in iowa has surrendered his license to practice and admitted to swapping services for sex and performing exorcisms on some patients." relevant to the reference? **Yes** Is this summary sentence 2 "Manuel also recommended that patients stop taking medication no longer exist before he can resume practicing chiropractic in the state." relevant to the reference? **Yes** Is this summary sentence 3 "The disgraced chiropractor received a perfect five out of five stars in patient satisfaction." relevant to the reference? **No**

Table 12: Case study on the evaluation of relevance in the SummEval dataset. The content in the bracket indicates the scale of evaluation scores in each metric, where higher scores mean better quality. The evidence of DecompEval denotes the subquestions with their answers.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7 (after the section of conclusions)

☒ A2. Did you discuss any potential risks of your work?
Our work cannot produce new contents. Our main goal is to build a state-of-the-art evaluation metric for text generation which shows great generalization ability and interpretability.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4

☑ B1. Did you cite the creators of artifacts you used?
4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix D.1

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We do not create new datasets. We follow many existing works to use benchmark datasets for fair comparison.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4.1, 4.5.2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4.1, 4.5.2

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4.2, Appendix D.3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Our proposed method is unsupervised which does not need hyperparameter search. We only need to follow existing works to set the input length for fair comparison.

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Our proposed unsupervised method can directly provide deterministic experimental results in a run.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix D.2

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
4.6

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix D.4

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix D.4

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix D.4

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We recruit annotators to measure the quality of texts in benchmark datasets, which is regarded as human labels to test the model performance.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We recruit three graduate students as annotators without collecting demographic and geographic information.