Towards a Unified Multi-Dimensional Evaluator for Text Generation

Ming Zhong[§] **Yang Liu**[†] **Da Yin**[♣] **Yuning Mao**[§] **Yizhu Jiao**[§]
Pengfei Liu[‡] **Chenguang Zhu**[†] **Heng Ji**[§] **Jiawei Han**[§]
[§]University of Illinois at Urbana-Champaign [†]Microsoft Cognitive Services Research
[♣]University of California, Los Angeles [‡]Carnegie Mellon University
{mingz5, yuningm2, yizhuj2, hengji, hanj}@illinois.edu
{yaliu10, chezhu}@microsoft.com da.yin@cs.ucla.edu pliu3@cs.cmu.edu

Abstract

Multi-dimensional evaluation is the dominant paradigm for human evaluation in Natural Language Generation (NLG), i.e., evaluating the generated text from multiple explainable dimensions, such as coherence and fluency. However, automatic evaluation in NLG is still dominated by similarity-based metrics, and we lack a reliable framework for a more comprehensive evaluation of advanced models. In this paper, we propose a unified multi-dimensional evaluator UniEval for NLG. We re-frame NLG evaluation as a Boolean Question Answering (QA) task, and by guiding the model with different questions, we can use one evaluator to evaluate from multiple dimensions. Furthermore, thanks to the unified Boolean QA format, we are able to introduce an intermediate learning phase that enables UniEval to incorporate external knowledge from multiple related tasks and gain further improvement. Experiments on three typical NLG tasks show that UniEval correlates substantially better with human judgments than existing metrics. Specifically, compared to the top-performing unified evaluators, UniEval achieves a 23% higher correlation on text summarization, and over 43% on dialogue response generation. Also, UniEval demonstrates a strong zero-shot learning ability for unseen evaluation dimensions and tasks. Source code, data and all pre-trained evaluators are available on our GitHub repository[1].

1 Introduction

The rapid development of Natural Language Generation (NLG) tasks with the support of pre-trained language models (Raffel et al., 2020; Brown et al., 2020; Lewis et al., 2020) calls for a higher quality evaluation of generated texts. However, the evaluation process is still dominated by traditional similarity-based metrics (Kasai et al., 2021), exemplified by ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) that compute n-gram overlap between the model output and the reference text. These metrics are potentially misleading as NLG models have advanced to the point where discrepancies between them are unlikely to be detected based on surface-level features (Gehrmann et al., 2022). Although using pre-trained models to obtain embedding-based similarity may alleviate this issue (Zhang et al., 2019), these metrics still naturally lead to the question: *does similarity to reference text indicate the overall quality of model output?* Belz and Gatt (2008) referred to this similarity as "human-likeness" and pointed out that the ability to output human-like text may be completely unrelated to the final performance on generation tasks.

Realizing that creating a one-size-fits-all score is infeasible, subsequent research has focused on a more comprehensive multi-dimensional evaluation for NLG tasks. It aims to evaluate the model output from multiple explainable dimensions and has been the dominant paradigm in human evaluation (Fabbri et al., 2021). For example, text summarization typically uses four dimensions for evaluation: `coherence`, `consistency`, `fluency`, and `relevance` (see Table 1). One way to achieve this fine-grained evaluation is to develop multiple evaluators dedicated to every single dimension (Dziri et al., 2019; Kryściński et al., 2020). However, it requires extensive effort to individually select and train an evaluator for each dimension when conducting multi-dimensional evaluations. On the other hand, several studies worked on building a unified evaluator, i.e., a single model that can produce multiple metrics (e.g., precision and recall) for the generated text (Yuan et al., 2021). Nevertheless, their evaluation scores cannot be directly aligned with the dimensions designed in human evaluation (e.g., `consistency` and `coherence`).

In this paper, we propose a unified multi-dimensional evaluator UniEval for text generation tasks. UniEval unifies all evaluation dimensions into a Boolean Question Answering (QA)

[1] https://github.com/maszhongming/UniEval

2023
Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2023 - 2038
December 7-11, 2022

Generated Summary: Harry Kane is nominated for both the PFA player and young player of the season. The Spurs striker has been released from the awards ceremony on Sunday. The Tottenham striker features in a new animation. **Reference Summary**: Harry Kane has been in superb form for Tottenham this season. The 21-year-old has scored 30 goals in all competitions for Spurs. Kane also made his England debut and scored within two minutes. **Document**: Harry Kane's celebrations this season have always shown him to be an animated young man ...			
Similarity-based Evaluators			
ROUGE-1: 0.44	ROUGE-2: 0.25	ROUGE-L: 0.42	BERTScore: 0.24
Single-dimensional Evaluators (predicted by two different evaluators (Deng et al., 2021))			
Consistency: 0.87	Relevance: 0.74		
Unified Evaluator (predicted by BARTScore, and the scoring range is negative infinity to 0)			
Precision: -5.45	Recall: -4.93	F_1: -5.19	
Multi-dimensional Evaluator (predicted by our evaluator UNIEVAL)			
Coherence: 0.04	Consistency: 0.41	Fluency: 0.92	Relevance: 0.28
Human Evaluation (annotated by experts, and we map the scoring range to 0-1)			
Coherence: 0.00	Consistency: 0.25	Fluency: 1.00	Relevance: 0.33

Table 1: An example of evaluating the text summarization task. All metrics except BARTScore are scored in the range of 0 to 1, with higher scores indicating better quality. Our proposed UNIEVAL is consistent with human evaluation, using multiple dimensions: `Coherence`, `Consistency`, `Fluency`, and `Relevance` to evaluate the generated text. The scores predicted by UNIEVAL are closer to human judgements.

problem (Clark et al., 2019a), thus enabling the evaluation of the generated text from different perspectives using only a single model. For instance, UNIEVAL can evaluate `coherence` in summarization by inputting a specific question, such as *"Is this a coherent summary to the document?"*. Moreover, thanks to the unified Boolean QA format, we are able to perform an intermediate training stage on four types of tasks related to NLG evaluation. This can be crucial for evaluation quality, since we lack large-scale human scores of model outputs to train an evaluator, a unified format that encompasses diverse existing tasks (namely, intermediate tasks) can substantially help UNIEVAL incorporate external knowledge related to NLG evaluations.

Specifically, a unified framework can bring the following benefits:

1) **Ease of use**. One model is sufficient, without the effort of picking multiple appropriate single-dimensional evaluators for all the dimensions.

2) **Internal complementarity**. Different dimensions in the same NLG task can be closely related to each other, so it is useful to perform joint training for these dimensions to share knowledge.

3) **External knowledge incorporation**. The unified Boolean QA format makes it possible to enhance the pre-trained language model by multi-task learning on diverse and relevant intermediate tasks before being trained on evaluation tasks.

4) **Extensibility and transferability**. A unified evaluator can achieve better extensibility and transferability with continual learning (Parisi et al., 2019) or prompting (Liu et al., 2021b; Chen et al., 2022), as it can accommodate more evaluation dimensions by modifying the input question.

Experimentally, UNIEVAL surpasses advanced evaluators by a large margin when evaluating three typical NLG tasks. Concretely, compared to the best unified evaluators (Yuan et al., 2021; Mehri and Eskenazi, 2020), UNIEVAL improves the correlation with human judgment by 23% on text summarization, and the improvement exceeds 43% on dialogue response generation. Ablation studies verify the effectiveness of our intermediate tasks. We also conduct transfer experiments and show that UNIEVAL achieves better performance compared with strong baseline metrics on unseen dimensions and NLG tasks in a zero-shot setting.

2 Related Work

Similarity-based Metrics Similarity-based metrics refer to the scores for evaluating the NLG models by measuring the similarity between a generated text and a reference text. They can be divided into lexical overlap-based (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005) as well as contextualized embedding-based (Zhang et al., 2019; Zhao et al., 2019; Clark et al., 2019b) evaluators. Although more than 60% of recent NLG papers solely use ROUGE or BLEU as the evaluation metric (Kasai et al., 2021), they fail to measure content quality (Reiter and Belz, 2009) and syntactic correctness (Stent et al., 2005), and are thus insufficient to portray the reliability of NLG systems.

Single-dimensional Evaluator To conduct more fine-grained evaluations for NLG, recent studies develop evaluators for a specific dimension, such as `consistency` in summarization (Kryściński et al., 2020; Wang et al., 2020; Cao et al., 2020; Durmus et al., 2020) and `coherence` in dialogue response generation (Dziri et al., 2019; Huang et al., 2020; Ye et al., 2021). These evaluators can help us better understand the characteristics of advanced NLG models from different perspectives. However, considering that most dimensions currently have no corresponding standard evaluators, solely using multiple single-dimensional evaluators to perform multi-dimensional evaluation is hard to achieve.

Unified Evaluator Several recent evaluators can predict multiple numbers for evaluating text by using different input and output contents (Yuan et al., 2021), multiple model variants (Mehri and Eskenazi, 2020), or different formulas (Scialom et al., 2021), and we refer to them as unified evaluators. These evaluation scores usually have no corresponding explanations or are simply categorized as precision, recall, and F_1, which poses difficulties in how to use them. Therefore, we propose a unified multi-dimensional evaluator in this paper, which attempts to align the evaluation scores with different dimensions in human evaluation.

3 Method

In this section, we first introduce how to formulate multi-dimensional evaluation as a unified Boolean QA problem, and then describe in detail the training paradigm for UniEval.

3.1 Problem Formulation

Multi-dimensional evaluation of NLG requires to evaluate n particular dimensions $d = (d_1, \ldots, d_n)$ of the model output, and the input can include the candidate output x, reference text y, and context c. y is removed when evaluating reference-independent dimensions, such as `consistency` in summarization. Depending on the specific generation task, c can contain different content or even be omitted. Evaluators need to evaluate the quality of the model output on each dimension and output scores $s = (s_1, \ldots, s_n)$ for all the dimensions.

To unify all evaluation dimensions into one evaluator, we transform each dimension into a Boolean question q_i. For example, for d_i = `coherence` in summarization, the transformed question q_i is *"Is this a coherent summary to the document?"*. Then for each input (x, y, c, q_i), evaluator should output "Yes" or "No" and calculate s_i as:

$$s_i = \frac{P(\text{"Yes"}|x, y, c, q_i)}{P(\text{"Yes"}|x, y, c, q_i) + P(\text{"No"}|x, y, c, q_i)}, \tag{1}$$

where $P(\cdot)$ denotes the probability of the model generating a specific word. In this way, a single evaluator can evaluate x of all dimensions by modifying the question description.

3.2 Unsupervised Learning on Multiple Evaluation Dimensions

Since annotating large-scale human scores to judge the quality of the generated text is unaffordable, we adopt an unsupervised setting to develop our evaluator. Using T5 (Raffel et al., 2020) as the backbone model, we first design specific rules for several commonly evaluated dimensions to construct pseudo data, and then combine them to train the evaluator.

Pseudo Data Construction To train an evaluator, we need to construct positive and negative samples for different dimensions. The former implies high-quality generated text, so we use groundtruth such as the reference summary in summarization. Then we propose particular rules for each dimension to convert positive samples into negative ones.

Taking text summarization as an example, the specific rule-based transformations are as follows: 1) **Coherence** refers to whether all the sentences form a coherent body. To build incoherent summaries, we use BM25 (Robertson and Zaragoza, 2009) to retrieve similar summaries, and randomly select a sentence from the retrieved summary to replace one of the sentences in groundtruth summary. 2) **Consistency** is the factual alignment between the summary and the source document. We use the method in Chen et al. (2021) to construct inconsistent summaries by antonym substitution, numerical editing, entity replacement, and syntactic pruning. 3) **Fluency** represents the quality of individual sentences. We randomly draw a span[2] from the positive sample and perform one of repeating, deleting, and shuffling to obtain the disfluent summaries. 4) **Relevance** means whether the summary contains only the important information of the source

[2]The length of the span is sampled from the Poisson distribution ($\lambda = 5$).

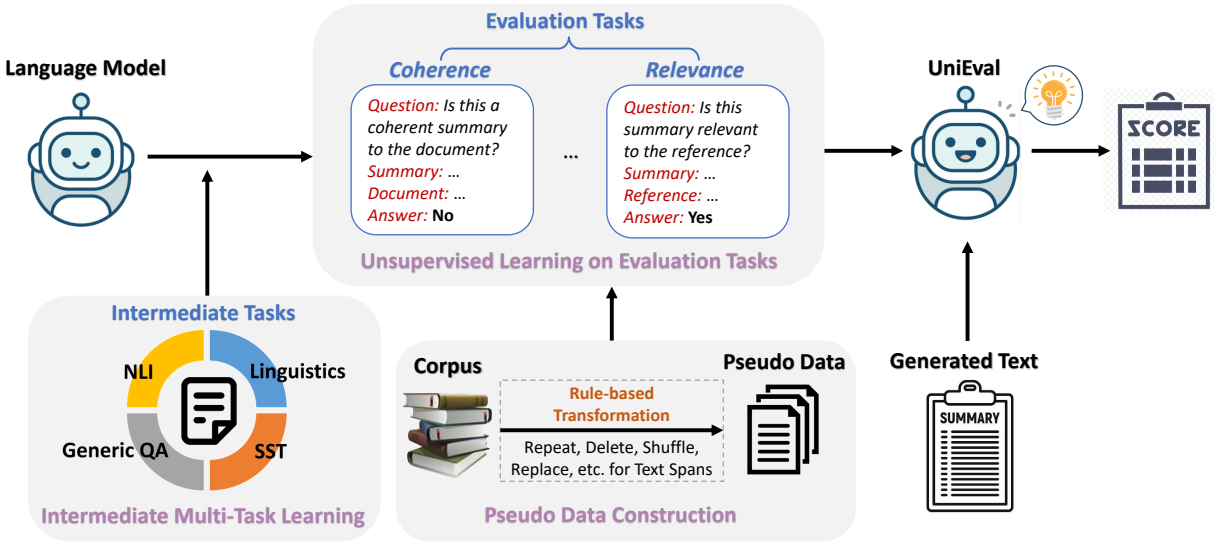

Figure 1: The overall framework of UNIEVAL. We convert all tasks into a Boolean QA problem and utilize the model to answer with "Yes" or "No". This unified QA format allows the model to incorporate external knowledge from multiple related tasks, i.e., intermediate multi-task learning. Then we construct pseudo data for each dimension and train them sequentially to obtain UNIEVAL.

document. The transformation rule is similar to coherence, except that we replace multiple sentences at random instead of one.

We include the designed rules for other NLG tasks in Appendix A.2. The detailed descriptions and concrete examples for all dimensions can also be found in Appendices A.1 and A.3.

Training Strategy For each generation task, we attempt to build a single evaluator to evaluate the NLG model from different dimensions. A straightforward approach is to perform multi-task learning on synthetic data of all dimensions to obtain a unified evaluator. However, we observe the negative transfer problem in several dimensions (e.g., `coherence` in summarization and `engagingness` in dialogue generation, see Tables 3 and 4). To tackle this issue, we employ a simple and effective method from continual learning (Parisi et al., 2019): whenever a new dimension is introduced, we add small portion of data from all previous dimensions to replay. The benefit is that we can easily extend our evaluator to new dimensions without training from scratch. Moreover, this method enables to explicitly learn dimensions related to linguistic features (e.g., `fluency`) first, and then move on to the dimensions that require a better understanding of the text (e.g., `consistency`). We show that this sequential training approach can alleviate the negative transfer problem in Section 4.

3.3 Intermediate Multi-task Learning

Benefiting from the unified Boolean QA format, we can additionally introduce intermediate tasks for UNIEVAL to incorporate external knowledge from existing related datasets. As shown in Figure 1, this stage is placed before the unsupervised learning on evaluation tasks. Notably, the input here is (c, q), which no longer includes the candidate output x and the reference text y. In total, we collect four types of intermediate tasks as follows.

Natural Language Inference. The task of NLI is to determine whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral) under a "premise". We transform the NLI task into a question: "*Is this hypothesis entailed in the premise?*", and only convert entailment into the label "Yes" and the rest to "No". The context c consists of a hypothesis and a premise. We use the following three datasets: document-level NLI (Yin et al., 2021), MRPC corpus (Dolan and Brockett, 2005) and QQP (Wang et al., 2017).

Self-Supervised Task. Based on the classical next sentence prediction task (Devlin et al., 2019), we propose a new *opening sentence prediction task*. The goal of this task is to determine whether a sentence is the starting sentence of a given news article. The motivation is that the first few sentences in the news tend to be salient and informative (See et al., 2017; Zhong et al., 2019), so it allows the model to learn inter-sentence coherence while also captur-

Tasks	# Positive	# Negative	Total
NLI	41,149	44,652	85,801
Self-supervised	30,000	30,000	60,000
Linguistics	6,744	2,850	9,594
Generic QA	17,032	13,096	30,128
All	94,925	90,598	185,523

Table 2: Statistics for intermediate tasks. Positive sample indicates the model should answer with "Yes".

ing the central idea of the document. We sample news from the CNN/DailyMail news corpus (Hermann et al., 2015) and randomly select the opening sentence of other news as negative samples.

Linguistics-Related Task. To facilitate the incorporation of linguistic knowledge into the unified model, we also include CoLA dataset (Warstadt et al., 2019) as the linguistic task. This requires the model to judge whether a sentence is linguistically acceptable, so the input question is: "*Is this a fluent and linguistically acceptable sentence?*".

Generic QA. We collect the existing Boolean QA datasets: BoolQ (Clark et al., 2019a), BoolQ-NP (Khashabi et al., 2020), BoolQ-CS (Gardner et al., 2020), StrategyQA (Geva et al., 2021), and extract the questions in MultiRC dataset (Khashabi et al., 2018) that can be answered with Yes/No as the data for generic QA task. Introducing these diverse question descriptions enables the model to better understand the importance of *question* in the input format as well as incorporate more open-ended external knowledge.

The statistics of data can be found in Table 2 and concrete examples for each task are also provided in Appendix B. Since this phase is not related to the evaluation metric, we train the model with cross-entropy loss without computing s_i.

4 Experiments

Following Deng et al. (2021), we classify NLG tasks into three types: compression, creation, and transduction, and select typical tasks from each category to conduct experiments. For compression and creation, we choose summarization and dialogue response generation to measure the performance of UNIEVAL, as well as the ability to zero-shot to unseen dimensions. For transduction, we select data-to-text to test whether UNIEVAL has the ability to transfer to a new NLG task.

4.1 Implementation Details

We use "google/t5-v1_1-large" version of T5 as the backbone model in all the experiments. The number of pseudo samples for each dimension is 30k, with an equal number of positive and negative examples. The order for continual learning is `coherence` → `fluency` → `consistency` → `relevance` for summarization, and `coherence` → `naturalness` → `groundedness` → `engagingness` for dialogue generation. For the score calculation, we follow previous work to compute sentence-level average scores for `fluency` and `consistency` (Laban et al., 2021) in summarization, and sentence-level cumulative scores for `engagingness` (Deng et al., 2021), while the rest is calculated as Equation 1. More details can be found in Appendix C.

4.2 Baselines

We compare UNIEVAL with several state-of-the-art evaluators. Notably, all the single-dimensional and unified evaluators are built on the same corpus.

BERTScore (Zhang et al., 2019) is a similarity-based evaluator. It computes the similarity between two text sequences based on the contextualized embedding obtained by BERT (Devlin et al., 2019).

MoverScore (Zhao et al., 2019) adds many-to-one alignment to BERTScore and introduces new aggregation methods to achieve a more powerful similarity-based evaluator.

CTC (Deng et al., 2021) utilizes information alignment to define metrics for several specific dimensions in NLG tasks, and proposes three model variants for each dimension. We compare the best variants of CTC in each dimension as the single-dimensional evaluators in our experiments.

BARTScore (Yuan et al., 2021) is a unified evaluator which uses average likelihood of the model output as the metric. It can predict different scores depending on the different input and output. We follow the original paper using $c \rightarrow x$ as the score for `coherence`, `consistency` and `fluency`, and $x \rightarrow y$ as the score for `relevance`.

USR (Mehri and Eskenazi, 2020) is a unified evaluator designed for dialogue response generation task. It uses different variants (e.g., MLM, dialogue retrieval and overall metric) to predict multiple scores for each generated response. We choose the score with the best correlation for each dimension for comparison in the experiments.

Metrics	Coherence		Consistency		Fluency		Relevance		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Similarity-based Metrics										
ROUGE-1	0.167	0.126	0.160	0.130	0.115	0.094	0.326	0.252	0.192	0.150
ROUGE-2	0.184	0.139	0.187	0.155	0.159	0.128	0.290	0.219	0.205	0.161
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
BERTSCORE	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
MOVERSCORE	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
Single-dimensional Evaluators										
CTC (Consistency)	0.223	0.172	0.415	0.345	0.335	0.276	0.166	0.124	0.285	0.229
CTC (Relevance)	0.402	0.310	0.366	0.301	0.299	0.245	0.428	0.336	0.374	0.298
UNIEVAL (Coherence)	**0.546**	**0.422**	0.337	0.280	0.324	0.266	0.418	0.316	0.406	0.321
UNIEVAL (Consistency)	0.176	0.127	**0.472**	**0.393**	0.366	0.300	0.176	0.128	0.298	0.237
UNIEVAL (Fluency)	0.324	0.247	0.276	0.229	**0.433**	**0.360**	0.236	0.176	0.317	0.253
UNIEVAL (Relevance)	0.543	0.420	0.324	0.267	0.340	0.283	**0.463**	**0.355**	0.417	0.332
Unified Evaluators										
BARTSCORE	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
UNIEVAL (Multi-task)	0.495	0.374	0.435	0.365	0.419	0.346	0.424	**0.327**	0.443	0.353
UNIEVAL (Continual)	**0.575**	**0.442**	**0.446**	**0.371**	**0.449**	**0.371**	**0.426**	0.325	**0.474**	**0.377**
- Intermediate Tasks	0.477	0.363	0.403	0.333	0.414	0.342	0.395	0.301	0.422	0.335

Table 3: Summary-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on SummEval benchmark. The underlined numbers indicate the results of transferring a single-dimensional evaluator to other dimensions.

4.3 Benchmarks

We adopt four meta-evaluation benchmarks for various NLG tasks to measure the correlation between UNIEVAL and human judgments.

SummEval (Fabbri et al., 2021) is a meta-evaluation benchmark for summarization. For each summary to be evaluated, it provides human scores from four dimensions: `fluency`, `coherence`, `consistency` and `relevance`. We use it to measure the performance of UNIEVAL.

Topical-Chat (Mehri and Eskenazi, 2020) is a benchmark for knowledge-based dialogue response generation task. It includes human scores from five dimensions: `naturalness`, `coherence`, `engagingness`, `groundedness` and `understandability`[3]. The first four dimensions are used to measure the performance of UNIEVAL, and the last one is used for the transfer experiment.

SFRES and **SFHOT** (Wen et al., 2015) are meta-evaluation benchmarks for data-to-text task. They provide information about restaurants and hotels in San Francisco and aim to let the model generate corresponding utterances. We leverage the annotations of `informativeness` and `naturalness` dimension to conduct transfer experiment.

QAGS (Wang et al., 2020) is also a benchmark for summarization. It is designed to detect `consistency` dimension on two summarization corpora (Narayan et al., 2018). We use it to test the performance of the single-dimensional version of UNIEVAL, and the results are listed in Appendix D.

4.4 Results For Summarization

Following Liu et al. (2021a), we use summary-level Spearman and Kendall-Tau correlation to assess the performance of different evaluators for summarization. Results of similarity-based metrics are listed in the first part of Table 3. They are designed to measure the semantic overlap between the model output and the reference text, so they can obtain relatively high correlations in `relevance` dimension. However, they are not qualified metrics for the other dimensions due to the poor correlation.

The second part contains the results of single-dimensional evaluators. CTC is currently the best evaluators of `consistency` and `relevance`, but it fails to excel on `coherence` and `fluency`. Here we also adapt UNIEVAL to several single-dimensional variants by training the model on pseudo data of only one dimension. Our proposed evaluators exceed CTC models and achieve the best correlation in all dimensions. It reveals that our proposed Boolean QA formulation can clearly enhance the backbone pre-trained model. Furthermore, we attempt to transfer the single-dimensional evaluators to other dimensions, and the underlined num-

[3]We rephrase (natural, maintains context, interesting, uses knowledge, understandable) from the original paper into the five dimensions mentioned above.

Metrics	Naturalness		Coherence		Engagingness		Groundedness		Average	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Similarity-based Metrics										
BLEU-1	0.161	0.133	0.210	0.223	0.314	0.334	0.289	0.303	0.243	0.248
BLEU-4	0.180	0.175	0.131	0.235	0.232	0.316	0.213	0.310	0.189	0.259
ROUGE-L	0.176	0.146	0.193	0.203	0.295	0.300	0.310	0.327	0.243	0.244
METEOR	0.212	0.191	0.250	0.302	0.367	0.439	0.333	0.391	0.290	0.331
BERTSCORE	0.226	0.209	0.214	0.233	0.317	0.335	0.291	0.317	0.262	0.273
Single-dimensional Evaluators										
CTC (Engagingness)	0.280	0.257	0.352	0.325	0.516	0.525	0.405	0.404	0.388	0.378
CTC (Groundedness)	0.200	0.161	0.256	0.228	0.485	0.475	0.524	0.477	0.366	0.335
UNIEVAL (Naturalness)	**0.500**	**0.547**	0.331	0.458	0.393	0.528	0.178	0.266	0.351	0.450
UNIEVAL (Coherence)	0.401	0.468	**0.543**	**0.607**	0.401	0.474	0.225	0.235	0.392	0.446
UNIEVAL (Engagingess)	0.394	0.427	0.471	0.477	**0.562**	**0.596**	0.376	0.431	0.451	0.483
UNIEVAL (Groundedness)	0.220	0.153	0.187	0.117	0.392	0.318	**0.543**	**0.511**	0.336	0.275
Unified Evaluators										
USR	0.337	0.325	0.416	0.377	0.456	0.465	0.222	0.447	0.358	0.403
UNIEVAL (Multi-task)	**0.480**	0.512	0.518	0.609	0.544	0.563	0.462	0.456	0.501	0.535
UNIEVAL (Continual)	0.444	**0.514**	**0.595**	**0.613**	**0.557**	**0.605**	**0.536**	**0.575**	**0.533**	**0.577**
- Intermediate Tasks	0.442	0.478	0.532	0.579	0.537	0.555	0.410	0.440	0.480	0.513

Table 4: Turn-level Pearson (r) an Spearman (ρ) correlations of different metrics on the Topical-Chat benchmark. The underlined numbers indicate the results of transferring a single-dimensional evaluator to other dimensions.

bers in Table 3 are transferred results. Overall UNIEVAL is better than CTC, but we can see that no single-dimensional evaluator can transfer well to all dimensions. For example, both `consistency` ⇒ `coherence` and `fluency` ⇒ `relevance` exhibit poor correlations, indicating that evaluators that focus solely on a single evaluation dimension lack acceptable transfer capability.

As shown in the last part, UNIEVAL substantially surpasses the state-of-the-art unified evaluator BARTScore in the summarization task. Specifically, UNIEVAL trained by multi-task learning brings an average improvement of more than 15% across all dimensions compared to BARTScore. And this gain is boosted to more than 23% by adapting continual learning in the unsupervised learning phase. The main gap between the two training strategies of UNIEVAL is the negative transfer on `coherence`, which clarifies that explicitly learning basic language features before learning more complex dimensions can alleviate this problem. It is also notable that compared with its single-dimensional version, the unified version of UNIEVAL is improved in both `coherence` and `fluency`, while having a slight decrement in the other two dimensions. This suggests that following continual learning, we can sequentially extend our evaluator to a new dimension while preserving the performance on previous dimensions. Moreover, the clear performance drop after removing the intermediate tasks in the last row illustrates the importance and usefulness of this phase.

4.5 Results For Dialogue Generation

To test the performance of different evaluators on the dialogue response generation task, we compute turn-level Pearson and Spearman correlation on the Topical-chat benchmark as in Mehri and Eskenazi (2020). Table 4 presents that similarity-based metrics correlate relatively well on `engagingness` and `groundedness` while performing poorly on the remaining dimensions. With respect to the single-dimensional evaluator, we can reach the same conclusion as for the summarization task: the scores predicted by UNIEVAL have the highest correlation with human judgments in all dimensions.

Compared to USR, the state-of-the-art unified evaluator in the dialogue response generation task, our evaluator demonstrates more remarkable boosts. According to Pearson and Spearman correlation, UNIEVAL (Continual) improves the results by an average of 48.9% and 43.2%, respectively. In comparison with the corresponding single-dimensional version, although there is a performance loss in `naturalness`, UNIEVAL (Continual) brings improvements in the remaining dimensions based on Spearman correlation. Especially for `grounedness`, the unified version increases the correlation by 12.5% (0.511 ⇒ 0.575) compared

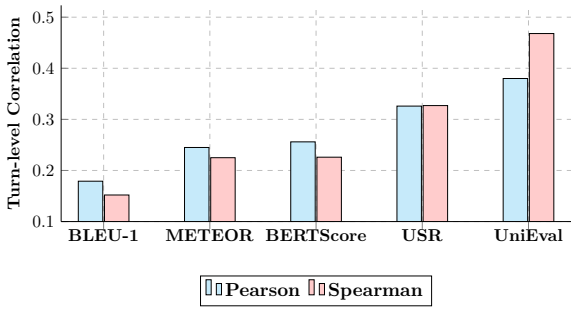

Figure 2: Zero-shot performance on the "understandability" dimension in dialogue response generation.

to the single-dimensional version. Meanwhile, intermediate tasks also display an indispensable role in evaluating dialogue generation, indicating that its benefits can span a variety of NLG tasks.

4.6 Transfer Experiments

We perform two zero-shot experiments to exhibit the transfer ability of UNIEVAL.

Zero-shot to Unseen Dimension To meet the requirements of different users, new evaluation dimensions often emerge for particular NLG tasks. For instance, certain users may prefer a new "`understandability`" dimension over other dimensions for the dialogue generation task. Therefore, we conduct experiments on the Topical-chat meta-evaluation benchmark to observe if UNIEVAL[4] has the transfer capability in this scenario. Concretely, we adjust the input question to *"Is this an understandable response in the dialogue?"*, and calculate the metric based on Equation 1. As shown in Figure 2, although UNIEVAL has not seen or been trained on this dimension before, its predicted score still correlates well with human judgments. It even outperforms the best USR metric for both Pearson (0.326 $\Rightarrow$ 0.380) and Spearman (0.327 $\Rightarrow$ 0.468) correlations, which denotes that UNIEVAL is capable of transferring to unseen dimensions by modifying the prompt.

Zero-shot to Unseen Task In a more radical setting, we also transfer UNIEVAL to a new NLG task of data-to-text generation in the zero-shot setting. As annotated in the SFRES and SFHOT benchmarks, users emphasize the `naturalness` and `informativeness` of the generated utterance for this task. Therefore, we adapt the question to "*Is this a fluent utterance?*" and "*Is this sentence informative according to the reference?*" to predict the evaluation scores for these two dimensions. "T5 + intermediate" in Table 5 represents the model obtained after the intermediate multi-learning stage. While it has not been trained on any evaluation tasks, it performs on par with BARTScore based on average correlations and is particularly good at evaluating the `naturalness` of utterances. After training on multiple evaluation dimensions on summarization, UNIEVAL (Summ)[5] demonstrates better transfer ability and superior performance over BARTScore in most dimensions of both datasets. This illustrates the capability of UNIEVAL to transfer to new NLG tasks without further adaptation.

[4]Here is UNIEVAL (Continual) in Table 4.

Metrics	SFRES		SFHOT		Avg.
	Nat.	Info.	Nat.	Info.	
ROUGE-1	0.170	0.115	0.196	0.118	0.150
ROUGE-L	0.169	0.103	0.186	0.110	0.142
BERTSCORE	0.219	0.156	0.178	0.135	0.172
MOVERSCORE	0.190	0.153	0.242	0.172	0.189
BARTSCORE	0.289	**0.238**	0.288	0.235	0.263
T5 + Intermediate	**0.348**	0.180	0.310	0.181	0.255
UNIEVAL (Summ)	0.333	0.225	**0.320**	**0.249**	**0.282**

Table 5: Zero-shot performance (Spearman) on the data-to-text task. Nat. and Info. denote `Naturalness` and `Informativeness`, respectively.

Evaluators	Coh.	Con.	Flu.	Rel.	Avg.
UNIEVAL	0.546	**0.472**	0.433	**0.463**	**0.479**
- NLI	0.532	0.417	**0.436**	0.452	0.459
- SST	0.498	0.462	0.428	0.450	0.460
- Linguistics	**0.548**	0.466	0.415	0.458	0.472
- Generic QA	0.528	0.438	0.421	0.436	0.456

Table 6: Ablation study of UNIEVAL. '-' means we remove this task from interemediate multi-task learning.

4.7 Ablation Study of Intermediate Tasks

We conduct ablation studies on the single-dimensional version of UNIEVAL to better investigate the contribution of each type of intermediate task to NLG evaluation. The results of Spearman correlation are presented in Table 6. Because of the similar task requirements, NLI contributes most to `consistency`, while our proposed opening sentence prediction task facilitates the evaluator to capture `coherence` between sentences. Due to the small data size of the linguistics-related task (see Table 2), removing it does not have a significant impact on the performance, but it can still help the

[5]Here is equivalent to UNIEVAL (Continual) in Table 3.

model better understand `fluency` of individual sentences. Generic QA enhances each dimension by engaging the evaluator to focus on the meaning of the input question. Overall, training on the combination of all four types of intermediate tasks leads to the best NLG evaluation performance.

5 Conclusion

In this paper, we emphasize the necessity of multi-dimensional evaluation in advancing the field of NLG. To promote this comprehensive and fine-grained evaluation approach, we propose a unified multi-dimensional evaluator UNIEVAL for various NLG tasks. UNIEVAL correlates well with human judgment on three typical generation tasks and exhibits excellent transfer performance.

Limitations

We state the limitations of this paper from the following four aspects:

1) Most of the current evaluators, including UNIEVAL, are black-box models. With the support of pre-trained language model, even though the neural evaluators can already correlate well with human judgments, it is still unclear how the model predicts these evaluation scores. Therefore, a better understanding of the evaluation process of different evaluators or the development of an interpretable and multi-dimensional evaluator may be the next stage for improving NLG evaluation.

2) Most of the neural evaluators are trained on synthetic data, while the pseudo data constructed in this paper still contain noise. For instance, for `fluency` in summarization, removing an unimportant span may not affect the fluency of the sentence, but we always treat the sentence after deleting as a negative sample. Thus, how to improve the quality of synthetic data could be an interesting topic.

3) We only use T5-large as the backbone model in the experiments due to the limited computational resources. How to extend the use of neural evaluators by using smaller models but retaining similar performance, or how to introduce more data to build larger evaluators with better performance, could be two future research directions.

4) We follow the categorization of NLG tasks in Deng et al. (2021) and select three typical tasks for our experiments, but UNIEVAL is still limited to English tasks. The generation tasks for cross-language scenarios are left for our future work.

Acknowledgements

We thank Weizhe Yuan, Mingkai Deng, Yu Meng, Hou Pong Chan, Dan Iter and Reid Pryzant for helpful discussions and feedback. We would also like to thank anonymous reviewers for valuable comments and suggestions. Research was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and INCAS Program No. HR001121C0165, National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Anja Belz and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200.

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.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *thirty-second AAAI conference on artificial intelligence*.

Yiran Chen, Pengfei Liu, and Xipeng Qiu. 2021. Are factuality checkers reliable? adversarial meta-evaluation of factuality in summarization. In *Find-*

ings of the Association for Computational Linguistics: EMNLP 2021, pages 2082–2095.

Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. Adaprompt: Adaptive model training for prompt-based nlp. *arXiv preprint arXiv:2202.04824*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019a. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019b. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3806–3812.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.

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*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. *Proc. Interspeech 2019*, pages 1891–1895.

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.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. Grade: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander R Fabbri, Yejin Choi, and Noah A Smith. 2021. Bidimensional leaderboards: Generate and evaluate language hand in hand. *arXiv preprint arXiv:2112.04139*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. More bang for your buck: Natural perturbation for robust question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 163–170.

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

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *arXiv preprint arXiv:2111.09525*.

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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

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. 2021a. Explainaboard: An explainable leaderboard for nlp. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

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.

GI Parisi, R Kemker, JL Part, C Kanan, and S Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks: the Official Journal of the International Neural Network Society*, 113:54–71.

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. *Journal of Machine Learning Research*, 21:1–67.

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

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Information Retrieval*, 3(4):333–389.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing*, pages 341–351.

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

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150.

Alex Warstadt, Amanpreet Singh, and Samuel Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve 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.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2718–2729.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. 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 (EMNLP-IJCNLP)*, pages 563–578.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058.

A Dimensions in Evaluation tasks

A.1 Explanation of Each Dimension

We introduce different dimensions for text summarization in Section 3.2. Here we include the detailed descriptions of different dimensions in dialogue response generation and data-to-text tasks.

For dialogue response generation (Mehri and Eskenazi, 2020):

- 1) `Naturalness`: judge whether a response is like something a person would naturally say
- 2) `Coherence`: determine whether this response serves as a valid continuation of the previous conversation.
- 3) `Engagingness`: determine if the response is interesting or dull.
- 4) `Groundedness`: given the fact that this response is conditioned on, determine whether this response uses that fact.
- 5) `Understandability`: judge whether the response is understandable.

For data-to-text (Wen et al., 2015):

- 1) `Naturalness`: determine whether the utterance could plausibly have been produced by a human.
- 2) `Informativeness`: determine whether the utterance contains all the information in the given content.

A.2 Pseudo Data Construction for Dialogue Response Generation

We produce pseudo data for the four dimensions of the dialogue response generation task as follows:

- 1) `Naturalness`: similar to `fluency` in summarization, except that we modify λ to 3.
- 2) `Coherence`: we randomly select gold response from other dialogues as the negative samples.
- 3) `Engagingness`: responses that are not engaging are dull and uninformative (Mehri and Eskenazi, 2020). So we let DialogGPT-small (Zhang et al., 2020) generate response given just one sentence, thus creating unattractive samples.
- 4) `Groundedness`: this dimension is used to measure how well the response refers to the knowledge context in knowledge-based conversations (Dinan et al., 2019). Therefore, we randomly extract a sentence from the current knowledge context and use a paraphrase generator[6] to rewrite it as a positive example, and sample a sentence from other knowledge contexts as a negative example.

A.3 Examples for Evaluation Tasks

We provide the concrete examples for different dimensions of evaluation tasks in Table 7. All the pseudo data is constructed on the CNN/DailyMail (Hermann et al., 2015) and Topical-Chat (Gopalakrishnan et al., 2019) corpus. We input reference text y (green text) to the model only when evaluating the `relevance` dimension in text summarization, while in the other dimensions UNIEVAL is a reference-free evaluator. Depending on the specific dimension, we feed the model with different contexts c. In addition, We use "\n" to separate the different turns in the dialogue history and end it with "\n\n".

B Examples for Intermediate Tasks

We also include the examples for each intermediate task in Table 8. We define the input as a (c, q) pair and let the model answer with "Yes" or "No".

C Implementation Details

We first train T5 on intermediate tasks for 2 epochs. For the evaluation tasks, we construct pseudo data on the CNN/DailyMail (Hermann et al., 2015) and Topical-Chat corpus (Gopalakrishnan et al., 2019) for summarization and dialogue generation, respectively. The number of samples for each dimension is 30k, with an equal number of positive and negative examples. We set batch size to 36 and the maximum learning rate to 5e-5 for both stages. Regarding continuous learning, we randomly select 20% of the data from the previously learned tasks to replay. The order is `coherence` → `fluency` → `consistency` → `relevance` for summarization, and `coherence` → `naturalness` → `groundedness` → `engagingness` for dialogue generation. Considering the difference in learning difficulty, we train 0.2-2.0 epochs for each dimension. And for multi-task learning in the evaluation

[6] huggingface.co/tuner007/pegasus_paraphrase

Dimensions	Tasks	Input	Target
Coherence	Summarization	question: Is this a coherent summary to the document? </s> summary: Theodore Wafer's statement contradicts his attorney's claim that he feared for his life and acted in self defense when he killed Renisha McBride. The shotgun is a Mossberg pump-action 12-gauge with a pistol grip. </s> document: On trial: Theodore Wafer, 55, initially told police that the shooting was an accident ...	Yes
Consistency	Summarization	question: Is this claim consistent with the document? </s> claim: The request wasn't rejected by Mr Justice Tugendhat last year and yesterday the Court of Appeal upheld that decision. </s> document: By James Slack The identity of suspects arrested by the police should be publicised before they are charged, the Court of Appeal has ruled ...	No
Fluency	Summarization	question: Is this a fluent paragraph? </s> paragraph: Jack Bowlby's body discovered at Cheltenham College at Cheltenham College . He was described as a "star pupil" by staff at the £30,000-a-year school.	No
Relevance	Summarization	question: Is this summary relevant to the reference? </s> summary: The Met Office issued severe weather warnings across the country yesterday as experts predicted a weekend washout. Forecasters said wind and rain will continue to batter the country until Tuesday at the earliest, as fears grew that the Somerset Levels could be flooded again. The North will be wet and windy for the next three days, with showers also scattered across the South West. </s> reference: Heavy rain caused massive tailbacks yesterday as a pothole opened up across three lanes of the M25. Fear grow that the Somerset Levels could be flooded again. North will be wet and windy for next three days, with showers also scattered across the South West.	Yes
Naturalness	Dialogue	question: Is this a natural response in the dialogue? </s> response: yes and that launched the career of many people, the most notable being han solo. the acting in the first was the most notable being han solo. the acting atrocious but got better as more movies were made. all told a great movie.	No
Coherence	Dialogue	question: Is this a coherent response given the dialogue history? </s> response: wow that is a lot of money for a logo. did you know corproate sponsors pay $1.12 billion on the nba last year!? </s> dialogue history: hi! do you like basketball? \n yes, i am a big raptors fan. it's crazy how much companies are paying to put their logos on nba jerseys. \n i am sure it is extremely high to advertise on jerseys. do you know how much? \n different team to team but everyone seems to get them these days. i did see geico is paying 6.5 million per year for their patch on the wizards jersey! \n\n	Yes
Engagingness	Dialogue	question: Is this an engaging and informative response according to the dialogue history and fact? </s> response: i'm so glad i'm not the only one who thinks this. </s> dialogue history: do you follow american politics? \n some, i am not surprised that the first phone number in the white house was 1. lol \n it definitely helped people reach the white house the fastest. i am surprised they still use floppy disks for storage. \n\n </s> fact: president jimmy carter turned all white house thermostats down to 65 degrees during the winter of 1977. the very first phone number of the white house was "1". jimmy carter had solar panels installed on the white house...and ronald reagan had them removed. there is a replica of the white house in atlanta which was built as a private home you can mail a birth announcement to the white house and they'll send you a congratulations card back.	No
Groundedness	Dialogue	question: Does this response use knowledge from the fact? </s> response: batman's city of gotham is located in new jersey and neither batman nor the villain the joker refer to one another by name. </s> fact: there was a batman villain named condiment king and he was defeated by slipping on his own ketchup. adam west has a batman logo on one of his molars according to dc canon; batman's gotham city is located in new jersey in their face to face confrontations, neither batman nor joker refer to one another by name. weird al yankovich did voiceover work in the most recent dc animated film "batman vs robin".	Yes

Table 7: Examples of different dimensions in evaluation tasks. The red text indicates the question q, the green text denotes the candidate output x, the yellow text is the reference text y, and the blue text represents the context c.

Tasks	Datasets	Input	Target
NLI	DocNLI	question: Is this a claim consistent with the premise? </s> claim: The two victims were teenagers. </s> premise: 2 seriously wounded in Grand Crossing shooting Two men were seriously wounded in a shooting Thursday evening in the Grand Crossing neighborhood on the South Side. The men Ž2014 ages 28 and 39 Ž2014 were shot at by someone inside a vehicle that pulled up to them at 6:11 p.m. in the 7300 block of South Dante, Chicago Police said. The older man was shot in his face and taken to Northwestern Memorial Hospital in serious condition, police said. The younger man took himself to Jackson Park Hospital with a gunshot wound to his shoulder in serious condition.	No
NLI	MRPC	question: Is this sentence equivalent to the reference? </s> sentence: The DVD-CCA then appealed to the state Supreme Court. </s> reference: The DVD CCA appealed that decision to the U.S. Supreme Court.	Yes
NLI	QQP	question: Is the following question equivalent to the reference? </s> question: Do you need a passport to go to Jamaica from the United States? </s> reference: How can I move to Jamaica?	No
SST	CNN/DailMail	question: Is this sentence the coherent first sentence of the document? </s> sentence: Diego Maradona thinks Steven Gerrard and England's defence should be held responsible for the defeat to Uruguay that crushed their World Cup hopes. </s> document: Bomb disposal experts examined the mortars and confirmed that two contained white phosphorous, a police spokesman said. It was not the first time that Hamas has attempted to target Israel using mortars containing white phosphorous, the spokesman said. White phosphorus ignites and burns, creating white smoke when it is exposed to oxygen. Militaries use it as a smoke screen to protect troops during combat ...	No
Linguistics	CoLA	question: Is this a fluent and linguistically acceptable sentence? </s> sentence: Mary questioned Joe's desire to eat cabbage, but only after I had questioned Sally's desire to.	No
QA	BoolQ	question: is confectionary sugar the same as powdered sugar? </s> context: Powdered sugar, also called confectioners' sugar, icing sugar, and icing cake, is a finely ground sugar produced by milling granulated sugar into a powdered state. It usually contains a small amount of anti-caking agent to prevent clumping and improve flow. Although most often produced in a factory, powdered sugar can also be made by processing ordinary granulated sugar in a coffee grinder, or by crushing it by hand in a mortar and pestle.	Yes
QA	StrategyQA	question: Is a Boeing 737 cost covered by Wonder Woman (2017 film) box office receipts? </s> term: Wonder Woman (2017 film) </s> description of term: American superhero film directed by Patty Jenkins </s> facts: The average cost of a US Boeing 737 plane is 1.6 million dollars. Wonder Woman (2017 film) grossed over 800 million dollars at the box office.	Yes
QA	MultiRC	question: Did Susan's sick friend recover? </s> context: Sent 1: Susan wanted to have a birthday party. Sent 2: She called all of her friends. Sent 3: She has five friends. Sent 4: Her mom said that Susan can invite them all to the party. Sent 5: Her first friend could not go to the party because she was sick. Sent 6: Her second friend was going out of town. Sent 7: Her third friend was not so sure if her parents would let her. Sent 8: The fourth friend said maybe. Sent 9: The fifth friend could go to the party for sure. Sent 10: Susan was a little sad. Sent 11: On the day of the party, all five friends showed up. Sent 12: Each friend had a present for Susan. Sent 13: Susan was happy and sent each friend a thank you card the next week.	Yes

Table 8: Examples for different intermediate tasks. Since these tasks are not relevant to the evaluation, we recognize all parts except question q as context c.

Metrics	QAGS-CNN			QAGS-XSUM			Average		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
ROUGE-1	0.338	0.318	0.248	-0.008	-0.049	-0.040	0.165	0.134	0.104
ROUGE-2	0.459	0.418	0.333	0.097	0.083	0.068	0.278	0.250	0.200
ROUGE-L	0.357	0.324	0.254	0.024	-0.011	-0.009	0.190	0.156	0.122
BERTSCORE	0.576	0.505	0.399	0.024	0.008	0.006	0.300	0.256	0.202
MOVERSCORE	0.414	0.347	0.271	0.054	0.044	0.036	0.234	0.195	0.153
FACTCC	0.416	0.484	0.376	0.297	0.259	0.212	0.356	0.371	0.294
QAGS	0.545	-	-	0.175	-	-	0.375	-	-
BARTSCORE	**0.735**	**0.680**	**0.557**	0.184	0.159	0.130	0.459	0.420	0.343
CTC (Consistency)	0.619	0.564	0.450	0.309	0.295	0.242	0.464	0.430	0.346
UNIEVAL (Consistency)	0.682	0.662	0.532	**0.461**	**0.488**	**0.399**	**0.571**	**0.575**	**0.465**

Table 9: Pearson (r), Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on QAGS benchmark.

tasks, we train the evaluator for 1-3 epochs in different NLG tasks. We train UNIEVAL on two A6000 GPUs for a total of 5 hours. If the meta-evaluation benchmark contains multiple references, we only use the first one as input.

In addition, although we can compute the scores for all dimensions directly from Equation 1, we slightly modify the score calculation for several certain dimensions due to their characteristics. For example, for `fluency` and `consistency` in summarization, disfluency and inconsistency are usually detected using sentences as the basic unit (Fabbri et al., 2021; Laban et al., 2021), so we split the model output x into several sentences and calculate the score s_{ij} for j-th sentence as:

$$s_{ij} = \frac{P(\text{“Yes”}|x_j, y, c, q_i)}{P(\text{“Yes”}|x_j, y, c, q_i) + P(\text{“No”}|x_j, y, c, q_i)}. \quad (2)$$

Then the final score for x in these two dimensions is $s_i = \sum_{j=1}^{m} s_{ij}/m$, where m is the number of sentences in x. Another special dimension is `engagingness` in dialogue generation. Since it indicates the total volume of interesting facts presented in the response (Deng et al., 2021), we use the summation to compute it as $s_i = \sum_{j=1}^{m} s_{ij}$. Therefore, the scoring range for `engagingness` is $[0, +\infty)$, while all others are $[0, 1]$.

D Results on QAGS

Advanced NLG models suffer from the problem of generating text that is inconsistent with the source document (Cao et al., 2018), which has led recent research to develop evaluators for evaluating the `consistency` dimension in summarization (Kryściński et al., 2020; Wang et al., 2020; Cao et al., 2020; Durmus et al., 2020). Therefore, we particularly compare the single-dimensional version of UNIEVAL for `consistency` with the state-of-the-art factuality checkers.

We conduct experiments on the QAGS meta-evaluation benchmark, which contains two different summarizaion corpus: CNN/DailyMail (Hermann et al., 2015) and XSum (Narayan et al., 2018). As shown in Table 9, BARTScore performs best on the more extractive[7] part (QAGS-CNN), but shows poor correlation on the more abstractive[8] subset (QAGS-Xsum). UNIEVAL (Consistency) correlates well in both parts of the data, especially in the more challenging Xsum dataset, greatly outperforming all previous consistency detectors. On average, UNIEVAL (Consistency) outperforms the state-of-the-art evaluator CTC by more than 30% based on Spearman and Kendall-Tau correlations. Thus, a high-performance single-dimensional evaluators can also be developed under our proposed framework.

[7]The reference summaries of CNN/DailyMail datasets tend to be copied from the original text.

[8]The words in the summary in Xsum dataset often do not appear in the original text.