UMSE: Unified Multi-scenario Summarization Evaluation

Shen Gao[1*] **Zhitao Yao**[1*] **Chongyang Tao**[2] **Xiuying Chen**[3] **Pengjie Ren**[1]
Zhaochun Ren[1] **Zhumin Chen**[1†]
[1]Shandong University, Qingdao, China
[2]Microsoft Corporation, Beijing, China
[3]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
{shengao,renpengjie,zhaochun.ren,chenzhumin}@sdu.edu.cn, yaozhitao@mail.sdu.edu.cn
chotao@microsoft.com, xiuying.chen@kaust.edu.sa

Abstract

Summarization quality evaluation is a non-trivial task in text summarization. Contemporary methods can be mainly categorized into two scenarios: (1) *reference-based*: evaluating with human-labeled reference summary; (2) *reference-free*: evaluating the summary consistency of the document. Recent studies mainly focus on one of these scenarios and explore training neural models built on pre-trained language models (PLMs) to align with human criteria. However, the models from different scenarios are optimized individually, which may result in sub-optimal performance since they neglect the shared knowledge across different scenarios. Besides, designing individual models for each scenario caused inconvenience to the user. Inspired by this, we propose **U**nified **M**ulti-scenario **S**ummarization **E**valuation Model (UMSE). More specifically, we propose a perturbed prefix tuning method to share cross-scenario knowledge between scenarios and use a self-supervised training paradigm to optimize the model without extra human labeling. Our UMSE is the first unified summarization evaluation framework engaged with the ability to be used in three evaluation scenarios. Experimental results across three typical scenarios on the benchmark dataset SummEval indicate that our UMSE can achieve comparable performance with several existing strong methods which are specifically designed for each scenario.[1]

1 Introduction

Quantitatively evaluating the quality of generated summary is a non-trivial task that can measure the performance of the summarization system (Lin, 2004; Ng and Abrecht, 2015; Zhang et al., 2020; Scialom et al., 2021), and can also be used as a reward model to give an additional training signal for the summarization model (Wu and Hu, 2018; Narayan et al., 2018; Scialom et al., 2019; Gao et al., 2019a, 2020a). The dominant evaluation methods are traditional word-overlap-based metrics like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002). Although these metrics are very easy to use, they cannot evaluate semantic similarity. In recent years, many researchers focus on semantic-based evaluation tools (Ng and Abrecht, 2015; Zhang et al., 2020; Zhao et al., 2019). Different to traditional metrics which only use one score to measure the quality of the summary, Zhong et al. (2022) propose to evaluate the summary quality in several dimensions (*e.g.,* coherence, consistency, and fluency) by calculating the similarity between the generated summary and the human-annotated summary.

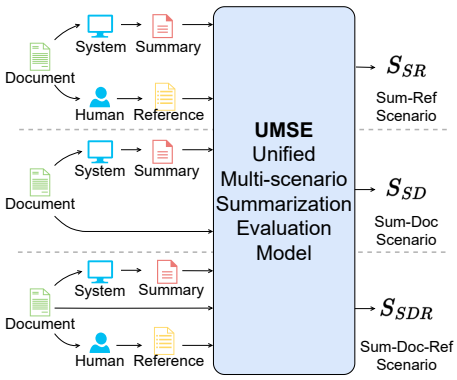

Figure 1: Illustration of multi-scenario summarization evaluation.

The summarization evaluation methods can be categorized into two scenarios based on the input data type: (1) **reference-based** methods require the human-annotated summary as input and (2) **reference-free** methods only use the corresponding document. The reference-based methods (Lin, 2004; Papineni et al., 2002; Banerjee and Lavie, 2005; Ng and Abrecht, 2015; Zhang et al., 2020; Zhao et al., 2019; Yuan et al., 2021) usually use the human written summary (*a.k.a.,*

[*] Equal contribution.
[†] Corresponding author.
[1] Code is available at https://github.com/ZT-Yao/UMSE.

Findings of the Association for Computational Linguistics: ACL 2023, pages 3837–3849
July 9-14, 2023

reference summary) as the ground truth and calculate the similarity between generated and reference summary. With the help of the pre-train language model, these methods have a powerful ability to measure semantic similarity. However, not all real-world application scenarios have human-annotated summaries. Using the reference-based evaluation method with the human-annotated ground truth summary is labor-consuming. Thus, reference-free methods (Wu et al., 2020; Gao et al., 2020b; Scialom et al., 2019, 2021) propose to evaluate the summary by modeling the semantic consistency between the generated summary and the document.

When evaluating a summarization system, even though we can individually select a proper evaluator condition on whether we have a reference summary, it is not very convenient. Moreover, since human annotation is costly, some summarization methods (Wu and Hu, 2018; Narayan et al., 2018; Scialom et al., 2019) choose to use the automatic evaluator to provide an additional training signal, instead of relying entirely on human-labeled document-summary pair data. In this type of usage, the evaluator needs to measure the quality of the model-generated summary with *partial* human-labeled document-summary data. Besides, contemporary trainable evaluation models for different scenarios (with or without reference summary) are built on pre-train language models, which may transfer knowledge across different scenarios and provides a great opportunity to bridge these evaluation scenarios with a better combination of the best of both worlds. Hence, it is valuable to build a unified multi-scenario summarization evaluator that can be used for processing both types of input data. Intuitively, this naturally leads to two questions: (1) *How to build a unified multi-scenario evaluation model regardless of whether we have a reference summary?* (2) *How to train the evaluator so that it can share knowledge between scenarios and maintain the exclusive knowledge in a specific task?*

In this paper, we propose a unified multi-scenario summarization evaluation method **U**nified **M**ulti-scenario **S**ummarization **E**valuation Model (UMSE). UMSE unifies three typical summary quality evaluation scenarios in one model: (1) **Sum-Ref**: evaluate using reference summary. UMSE measures the similarity between the generated summary and the human-annotated reference summary. (2) **Sum-Doc**: evaluate using document. Since using the reference summary is labor-consuming, UMSE can measure the consistency between generated summary and the original document. (3) **Sum-Doc-Ref**: evaluate using both document and reference summary. This method incorporates the advantages of sum-ref and sum-doc. To process these different types of input, we propose a perturbed prefix method based on the prefix tuning method (Li and Liang, 2021; Liu et al., 2022, 2021) that shares a unified pre-train language model across three scenarios by using different continuous prefix tokens as input to identify the scenario. Then, we propose 2 hard negative sampling strategies to construct a self-supervised dataset to train the UMSE without additional human annotation. Finally, we propose an ensemble paradigm to combine these scenarios into a unified user interface.

To sum up, our UMSE can bring the following benefits:

• **One model adaptable to multi-scenario**. UMSE uses only one model to evaluate the generated summary whenever it has a reference summary.

• **Mutually enhanced training**. We propose a perturbed prefix method to transfer knowledge between scenarios, and it can boost the performance of each scenario.

• **Self-supervised**. UMSE can be trained using a fully self-supervised paradigm without requiring any human-labeled data, and it makes UMSE has strong generalization ability.

To verify the effectiveness of the UMSE, we first compare with several baselines including the reference-based and reference-free methods. Specifically, UMSE outperforms all the strong reference-free evaluation methods by a large margin and achieves comparable performance with the state-of-the-art in a unified model. Ablation studies verify the effectiveness of our proposed perturbed prefix-tuning method.

2 Related Work

2.1 Reference-free Metrics

Reference-free metrics aim to evaluate the summary quality without the human-labeled ground truth summary as the reference, and these methods can be categorized into two types: trained model and training-free model. For the training-free methods, SUPERT (Gao et al., 2020b) first extracts salient sentences from the source document to construct the pseudo reference, then computes

the semantic similarity to get the evaluation score. Following SUPERT, Chen et al. (2021) propose a centrality-weighted relevance score and a self-referenced redundancy score. While computing the relevance score, the sentences of pseudo reference are weighted by centrality, the importance of each sentence. For the methods which should be trained, LS-Score (Wu et al., 2020) is an unsupervised contrastive learning framework consisting of a linguistic quality and a semantic informativeness evaluator. The question-answering paradigm is usually used in evaluating summaries, which evaluates the factual consistency between summary and document with the help of well-trained question-answering models (Scialom et al., 2019; Gao et al., 2019b; Durmus et al., 2020; Scialom et al., 2021).

2.2 Reference-based Metrics

Referenced-based metrics, which evaluate the quality of the summary by measuring the similarity of the summary and human written reference, can be divided into two categories: lexical overlap-based metrics and semantic-based metrics. ROUGE (Lin, 2004), the most commonly used metric for summary evaluation, measures the number of matching n-grams between the system output and reference summary. Other popular lexical overlap-based metrics are BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) which are also commonly employed in other text generation tasks (*e.g.,* machine translation). Since using the lexical overlap to measure the quality is sometimes too strict, many researchers turn to focus on exploring the semantic-based evaluation. ROUGE-WE (Ng and Abrecht, 2015) improves ROUGE by using Word2Vec (Mikolov et al., 2013) embeddings, and S3 (Peyrard et al., 2017) takes the ROUGE and ROUGE-WE as input features and is trained on human-annotated datasets. With the prosperity of the pre-training language model (PLM), more and more researchers introduce these models for evaluation. BERTScore (Zhang et al., 2020) leverages the contextual embeddings from BERT (Devlin et al., 2019) and calculates the cosine similarity between system output and reference sentence. CTC (Deng et al., 2021) is based on information alignment from two dimensions: consistency and relevance. UniEval (Zhong et al., 2022) is a multi-dimensional evaluator based on T5 (Raffel et al., 2020), and it formulates the summary evaluation as a binary question-answering task and evaluates from four dimensions: coherence, consistency, fluency, and relevance. However, existing summarization evaluation models usually focus on measuring the summary quality from multiple aspects and transferring knowledge from PLM, they ignore the shareable knowledge between different scenarios.

Evaluating the quality of the generated text is a also crucial task in generation tasks. In machine translation evaluation, Wan et al. (2022) proposes UniTE which is a multi-scenario evaluation method. UniTE employs monotonic regional attention to conduct cross-lingual semantic matching and proposes a translation-oriented synthetic training data construction method. However, the summarization task does not have these characteristics and directly applying UniTE to summarization evaluation cannot measure the important aspect of summary (*e.g.,* coherence and relevance).

3 UMSE Model

Problem Formulation Given a model-generated summary $X = \{x_1, x_2, \ldots, x_{L_x}\}$ with L_x tokens, our goal is to use a unified evaluation model to produce a score $s \in \mathcal{R}$ for X. For the Sum-Ref scenario, the model uses generated summary X and ground truth summary $Y = \{y_1, y_2, \ldots, y_{L_y}\}$ as input. For the Sum-Doc scenario, we evaluate the summary quality by using generated summary X and document $D = \{d_1, d_2, \ldots, d_{L_d}\}$ with L_d tokens as input, which does not require any human annotation (*e.g.,* ground truth summary Y). For the Sum-Doc-Ref scenario, the model uses generated summary X, ground truth summary Y, and document D as input. To train the evaluation model, we do not use any human-annotated summary quality dataset and we construct the training dataset by using several self-supervised training strategies.

3.1 Overview

In this section, we detail the **U**nified **M**ulti-scenario **S**ummarization **E**valuation Model (UMSE). An overview of UMSE is shown in Figure 2. UMSE has two main parts: (1) **Data construction.** We first construct two self-supervised datasets for coherence and relevance evaluation scenarios. (2) **Unified Model.** To unify the different input data into a unified model, we propose a perturbed prefix-tuning method to train the UMSE.

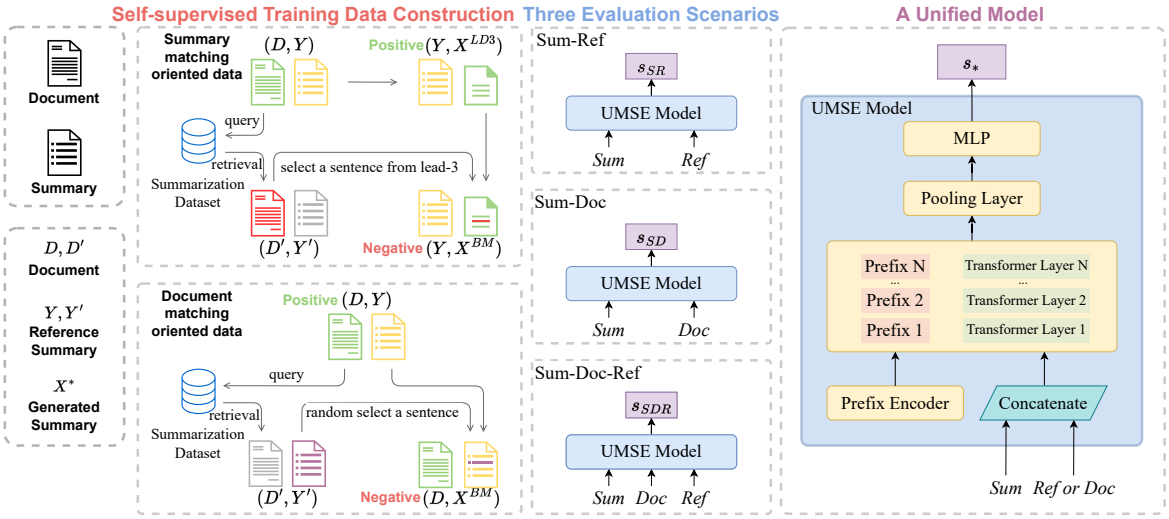

Figure 2: Illustration of UMSE which tackles the summarization evaluation in three scenarios by a unified model trained with two self-supervised tasks.

3.2 Data Construction

Employing a human annotator to annotate the quality of generated summary to train the evaluation model is labor-consuming and will lead the evaluation model hard to use. We propose to use the self-supervised tasks to construct the training dataset for the evaluator without using any human annotation. Since measuring the quality of the summary requires two main semantic matching abilities: (1) matching with the reference summary and (2) matching with the document, we propose two self-supervised tasks to construct the training dataset automatically:

• **Summary matching oriented data**: The goal for this task is to construct positive and negative samples which are different in whether the summary contains the salient information. Given a document-summary pair D, Y, the data sample to construct is a summary pair. The positive data pair (Y, X^{LD3}) contains the reference summary Y and a candidate summary X^{LD3} which contains relevant information. And the negative data pair (Y, X^{BM}) contains the reference summary Y and a candidate summary X^{BM} which describes similar but not relevant information. Particularly, if the negative data is very hard for the evaluation model to identify (*e.g.,* requires reasoning ability or is very similar to the positive sample), the evaluation model will achieve better performance than using very simple negative data. Thus, we propose to use the leading three sentences of the corresponding document D as the candidate summary X^{LD3}. For the candidate summary X^{BM} in negative data pair, we first use the BM25 retrieval model to retrieve the most similar document D' to D and obtain the reference summary Y' of D'. To make the negative sample harder, we randomly replace a sentence in Y' with one sentence in X^{LD3} as the final negative summary X^{BM}.

• **Document matching oriented data**: The golden criterion for evaluating the summary quality is whether the summary describes the main facts of the document. Hence, we construct self-supervised data which aims to train the model to measure the semantic relevance between summary and document. The positive data pair (D, Y) consists of document D and its reference summary Y. The negative data pair (D, X^{BM}) contains the document D and a false summary X^{BM} which is similar to Y. We employ the same BM25 retrieval method in coherence data construction to obtain Y' and replace a sentence in Y with a sentence in Y' as the negative summary X^{BM}.

For brevity, we omit the superscript of X in the following sections.

3.3 Perturbed Prefix-Tuning

Although the three scenarios have different input types, we can directly concatenate them into a text sequence which can be easily adopted by the pre-train language model. Following previous work (Zhong et al., 2022), although our evaluation model does not require additional summarization-quality data annotations, human-written summaries are still required to train the estimator. Therefore, reducing the dependence on human-written sum-

maries can improve the applicability of our model in low-resource scenarios. Thus, we employ prefix-tuning to explore the semantic understanding ability of large language models on the summarization evaluation task. Specifically, we append different prefix sequences at the start of each input text sequence according to the scenario:

$$\begin{aligned}\mathbf{H}_{SR} &= \text{PLM}([\text{CLS}]P_{SR}X[\text{SEP}]Y),\\ \mathbf{H}_{SD} &= \text{PLM}([\text{CLS}]P_{SD}X[\text{SEP}]D),\\ \mathbf{H}_{SDR} &= \text{PLM}([\text{CLS}]P_{SDR}X[\text{SEP}]D[\text{SEP}]Y),\end{aligned}$$

where [CLS] and [SEP] are both special tokens in PLM, $\mathbf{H}_{SR} \in \mathbb{R}^{(L_x+L_y+L_p+2),z}$ denotes the token level representation for Sum-Ref pair, and z is the hidden size of the PLM. The $\mathbf{P}_* \in \mathbb{R}^{\mathbf{L_p},\mathbf{z}}$ denotes the prefix for each scenario, which is a continuous prompt with length L_p. The advantage of using the unified evaluator is that we can use one large language model to conduct three tasks and it will reduce the size of the evaluation toolkit.

Although these data scenarios have their exclusive task characteristic, there are also some shared abilities and knowledge which can be transferred between different scenarios. To model the exclusive characteristic and transfer knowledge using the continuous prefix in a coordinated way, we propose a prefix perturbation method that uses the same tokens with different orders of different scenarios. Take the prefix of Sum-Doc scenario as an example, $\mathbf{P_{SD}}$ contains L_p continuous prefix tokens $\mathbf{P_{SD}} = \{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_{L_p}}\}$. We perturb $\mathbf{P_{SD}}$ as $\{\mathbf{p_1}, \mathbf{p_3}, \ldots, \mathbf{p_{L_p}}, \mathbf{p_2}, \ldots, \mathbf{p_{L_p-1}}\}$, and use this perturbed prefix as the prefix for Sum-Doc-Ref $\mathbf{P_{SDR}}$. This prefix perturbation method keeps the prefix used across scenarios to use the same continuous tokens in a different order. Thus, our model can simultaneously transfer knowledge between scenarios and keep the exclusive ability prompted by the different prefixes.

To obtain the summary-level overall representation, we conduct a pooling operation on the token-level representation:

$$\mathbf{E}_{SR} = \text{Pooling}(\mathbf{H}_{SR}), \quad (1)$$
$$\mathbf{E}_{SD} = \text{Pooling}(\mathbf{H}_{SD}), \quad (2)$$
$$\mathbf{E}_{SDR} = \text{Pooling}(\mathbf{H}_{SDR}), \quad (3)$$

where $\mathbb{E}_* \in \mathbb{R}^z$ denotes the summary-level representation. Then we employ a multi-layer perception (MLP) network to conduct a binary classification and obtain the probability p:

$$p_* = \text{Softmax}(\text{MLP}(\mathbf{E}_*)) \in \mathcal{R}^2, \quad (4)$$
$$s_* = p_*^+, \quad (5)$$

where $p_*^+ \in \mathbb{R}$ denotes the probability of positive class in p_*. During training, we use cross entropy loss $\mathcal{L}_{ce}$ to optimize the model parameters to distinguish the positive and negative samples:

$$\mathcal{L}_{ce} = -\left[\sum_{i=1}^{n} c_i \log p_i^+ + (1-c_i)\log\left(1-p_i^+\right)\right],$$

where $c_i \in \{0, 1\}$ denotes the label of i-th training sample which indicates whether this sample is a positive or negative sample. At the inference stage, we take the probability of positive class p^+ as the final evaluation score s.

3.4 Variant of Sum-Doc-Ref Evaluation

Intuitively, the scenario Sum-Doc-Ref can be seen as a combination of the Sum-Doc and Sum-Ref scenarios. Hence, an intuitive method to conduct the evaluation of the Sum-Doc-Ref scenario is to directly fuse the scores of the Sum-Doc and Sum-Ref scenarios. In this section, we propose a variant implementation to conduct evaluation conditions on the input of Sum-Doc-Ref, named **UMSE(Fusion)**. We combine the score of the Sum-Doc and Sum-Ref scenarios to get the score for the Sum-Doc-Ref:

$$s_{SDR} = f(s_{SR}, s_{SD}), \quad (6)$$

where f denotes the ensemble strategy, such as min and max. In the experiment, we will analyze the performance of different implementations of f.

4 Experiment

4.1 Datasets

In the training phase, we construct the positive and negative data pairs using the CNN/DailyMail (Nallapati et al., 2016) dataset. Then the trained evaluators are tested on the meta-evaluation benchmark SummEval (Fabbri et al., 2021) to measure the rank correlation coefficient between the evaluation model and human judgment.

CNN/DailyMail has $286,817$ training document-summary pairs, $13,368$ validation and $11,487$ test pairs in total. The documents in the training set have 766 words and 29.74 sentences on average while the reference summaries contain 53 words and 3.72 sentences.

SummEval is a meta-evaluation benchmark. To collect the human judgments towards the model-generated summaries, they first randomly select 100 document and reference pairs from the test set of CNN/DailyMail, then generate summaries using 16 neural summarization models. Each summary is annotated by 3 experts and 5 crowd-sourced workers along four dimensions: coherence, consistency, fluency, and relevance. Finally, there is a total of 12800 summary-level annotations.

4.2 Evaluation Metrics

Following previous work (Yuan et al., 2021; Zhong et al., 2022), we measure the rank correlation coefficient between the evaluation model and human judgment to represent the performance of the evaluator. In the experiments, we employ the Spearman (ρ) and Kendall-Tau (τ) correlations between the evaluator output scores and human ratings. The statistical significance of differences observed between the performance of UMSE and the strongest baseline in each scenario is tested using a two-tailed paired t-test and is denoted using $^{\blacktriangle}$ (or $^{\blacktriangledown}$) for strong significance at $\alpha = 0.01$ and $p < 0.05$.

4.3 Comparisons

In the experiment, we compare the proposed UMSE with widely used and strong baselines:

Reference-based Methods:

(1) ROUGE (Lin, 2004) is one of the most popular metrics, and it computes n-gram overlapping between the system output and reference summary. We employ the ROUGE-1, ROUGE-2, and ROUGE-L in our experiments. (2) BERTScore (Zhang et al., 2020) leverages the contextual embedding from the pre-training language model BERT (Devlin et al., 2019) and calculates the cosine similarity between system output and reference. (3) MoverScore (Zhao et al., 2019) utilizes the Word Mover's Distance to compute the distance between the embedding of generated summary and reference. (4) BARTScore (Yuan et al., 2021) uses the weighted log probability of the pre-train language model BART's (Lewis et al., 2020) output to evaluate the quality of summaries. (5) CTC (Deng et al., 2021) is a general evaluation framework for language generation tasks including compression, transduction, and creation tasks. CTC is designed on the concept of information alignment. (6) UniEval (Zhong et al., 2022) formulates the summary evaluation as binary question answering and can evaluate the summary from four dimensions, coherence, consistency, fluency, and relevance.

Reference-free Methods:

(1) BLANC (Vasilyev et al., 2020) is defined as a measure of the helpfulness of a summary to PLM while PLM performs the Cloze task on document sentences. In specific, the final score is the accuracy difference of whether use a summary to concatenate with the masked sentence. (2) SummaQA (Scialom et al., 2019) is a QA-based evaluation metric. It generates questions from documents, answers the questions based on the summary by a QA model, and computes the QA metric as evaluation scores. (3) SUPERT (Gao et al., 2020b) constructs the pseudo reference by extracting salient sentences from the source document and computes the similarity between generated summary and pseudo reference to evaluate the quality of the summary. (4) UniTE (Wan et al., 2022) is a unified evaluation model for machine translation in different scenarios: reference-only, source-only and source-reference-combined.

To prove the effectiveness of the perturbed prefix-tuning, we design an ablation model, UMSE-PT (w/o Prefix-Tuning). We remove the prefix of input and jointly fine-tune one pre-train language model using the two datasets we constructed.

4.4 Implementation Details

Following (Deng et al., 2021), we employ the roberta-large (Liu et al., 2019) as the backbone of our model. The MLP consists of 3 linear layers with tangent activation and the dimensions of each layer are 3072, 1024, and 2, respectively. Following (Wan et al., 2022), the max length of input sequence (with prompt) is set to 512. We vary the length of prompt in {8, 16, 32, 64, 128}, and find that 128 is the best choice. We use AdamW as the optimizer and the learning rate is set to 3.0e-05 selected from {2.0e-05, 3.0e-05, 5.0e-05}. The number of train epochs is set up to 10 epochs and the batch size is set to 8. We fix the random seed always to 12 and trained our model on an NVIDIA GeForce RTX 3090 GPU for 6-7 hours. We use PyLucene to implement the BM25 algorithm to retrieve similar documents. The size of the two training datasets is 30K respectively, and the positive and negative samples are half.

4.5 Evaluation Results

We compare our UMSE with strong baselines in Table 1. We can surprisingly find that UMSE (w/

Model	Coherence		Consistency		Fluency		Relevance	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Sum-Ref Methods								
ROUGE-1 (Lin, 2004)	0.1670	0.1260	0.1600	0.1300	0.1590	0.0940	0.3260	0.2520
ROUGE-2 (Lin, 2004)	0.1840	0.1390	0.1870	0.1550	0.1590	0.1280	0.2900	0.2190
ROUGE-L (Lin, 2004)	0.1280	0.0990	0.1150	0.0920	0.1050	0.0840	0.3110	0.2370
➤ BERTScore (Zhang et al., 2020)	0.2840	0.2110	0.1100	0.0900	0.1930	0.1580	0.3120	0.2430
MOVERScore (Zhao et al., 2019)	0.1590	0.1180	0.1570	0.1270	0.1290	0.1050	0.3180	0.2440
UniTE (w/ SR)(Wan et al., 2022)	0.1792	0.1362	0.0557	0.0474	0.0761	0.0614	0.2255	0.1716
UMSE (w/ SR)	0.5840▲	0.4443▲	0.2494▲	0.2055▲	0.2601▲	0.2132▲	0.4217▲	0.3189▲
UniEval (Zhong et al., 2022)	0.4950	0.3740	**0.4350**	**0.3650**	**0.4190**	**0.3460**	0.4240	0.3270
BARTScore (Yuan et al., 2021)	0.4480	0.3420	0.3820	0.3150	0.3560	0.2920	0.3560	0.2730
Sum-Doc Methods								
BLANC (Vasilyev et al., 2020)	0.1219	0.0951	0.2768	0.2307	0.1727	0.1436	0.2574	0.1983
SummaQA (Scialom et al., 2019)	0.1239	0.0963	0.2540	0.2102	0.1782	0.1457	0.2120	0.1628
➤ SUPERT (Gao et al., 2020b)	0.2165	0.1716	0.3438	0.2863	0.2509	0.2024	0.2746	0.2132
UniTE (w/ SD)(Wan et al., 2022)	0.1703	0.1327	0.1160	0.0956	0.0871	0.0703	0.2738	0.2084
UMSE (w/ SD)	0.5298▲	0.4052▲	0.3579▲	0.2961▲	0.3163▲	0.2617▲	0.4039▲	0.3060▲
Sum-Doc-Ref Methods								
➤ CTC (Deng et al., 2021)	0.4020	0.3100	0.3660	0.3010	0.2990	0.2450	0.4280	0.3360
UniTE (w/ SDR) (Wan et al., 2022)	0.1885	0.1453	0.1244	0.1017	0.1076	0.0886	0.2874	0.2232
UMSE (w/ SDR)	0.4704	0.3532	0.3413	0.2817	0.3006	0.2451	0.3894	0.2929
UMSE(Fusion) (w/ SDR)	**0.5944**▲	**0.4515**▲	0.3381	0.2813	0.3316▲	0.2731▲	**0.4358**▲	**0.3282**▲
Ablation Methods								
UMSE-PT (w/ SR)	0.5607	0.4246	0.2664	0.2193	0.2552	0.2079	0.4228	0.3155
UMSE-PT (w/ SD)	0.5007	0.3810	0.3505	0.2905	0.3079	0.2533	0.4276	0.3220
UMSE(Fusion)-PT (w/ SDR)	0.5757	0.4397	0.3338	0.2751	0.3206	0.2638	0.4375	0.3291

Table 1: Comparing with baselines on SummEval dataset. We use the notion "(w/ *)" to denote which data is used as input. (ρ) denotes the Spearman correlations and (τ) denotes the Kendall-Tau correlations. The row with shaded background denotes the multi-dimensional metrics which output a score for each dimension, and it is unfair for comparing with these methods. The number with underline denotes the max value in the scenario and the **bold-face** denotes the max value over three scenarios.

Fusion Methods	Coherence		Consistency		Fluency		Relevance	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Min	0.5896	0.4475	0.2729	0.2261	0.2786	0.2284	0.4315	0.3267
Max	0.5386	0.4099	**0.3561**	**0.2947**	0.3183	0.2624	0.4083	0.3084
Geometric mean	0.5938	0.4503	0.3151	0.2618	0.3132	0.2584	0.4332	0.3260
Arithmetic mean ✓	**0.5944**	**0.4515**	0.3381	0.2813	**0.3316**	**0.2731**	**0.4358**	**0.3282**

Table 2: Result of different fusion methods in Sum-Doc-Ref scenario.

	Coherence		Consistency		Fluency		Relevance	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Single Model (w/ SR)	0.5019	0.3796	0.2916	0.2391	0.3090	0.2525	0.4153	0.3096
UMSE (w/ SR)	0.5840↑	0.4443↑	0.2494↓	0.2055↓	0.2601↓	0.2132↓	0.4217↑	0.3189↑
Single Model (w/ SD)	0.4798	0.3599	0.3132	0.2580	0.2992	0.2454	0.3644	0.2760
UMSE (w/ SD)	0.5298↑	0.4052↑	0.3579↑	0.2961↑	0.3163↑	0.2617↑	0.4039↑	0.3060↑
Single Model (w/ SDR)	0.3488	0.2660	0.2824	0.2342	0.2739	0.2253	0.2435	0.1866
UMSE (w/ SDR)	0.4704↑	0.3532↑	0.3413↑	0.2817↑	0.3006↑	0.2451↑	0.3894↑	0.2929↑

Table 3: Comparison between UMSE and separately fine-tuning PLM.

SD) performs comparably to the UMSE (w/ SR) in the Sum-Ref scenario and achieves significant improvement over the existing baselines, which demonstrates that our proposed perturbed prefix-tuning can transfer knowledge from other scenarios. BERTScore is the state-of-the-art reference-based single-dimensional evaluation method, and the performance of UMSE increases by 105.63%, 34.93%, and 38.62% compared to BERTScore in terms of Coherence (ρ), Fluency (τ), and Relevance (ρ) respectively. Compared with the reference-free baselines, UMSE (w/ SD) outperforms SUPERT 144.71%, 29.30%, and 47.09% in terms of Coherence (ρ), Fluency (τ), and Relevance (ρ) respectively. Although the UMSE achieves slightly lower performance than the baseline in one dimension, the UMSE achieves consistently strong performance in three scenarios which can facilitate

Model	Faithful	Factual
ROUGE-1	0.197	0.125
ROUGE-2	0.162	0.095
ROUGE-L	0.162	0.113
BERTScore	0.190	0.116
QA	0.044	0.027
UMSE	**0.242**	**0.167**
Entailment	0.431	0.264

Table 4: The performance of different models on detecting hallucinations. The evaluation metric is the Spearman correlation. The faithful and factual annotations are released by Maynez et al. (2020). The row with shaded background denotes the model is trained on a supervised dataset, making it unfair to compare it with other methods.

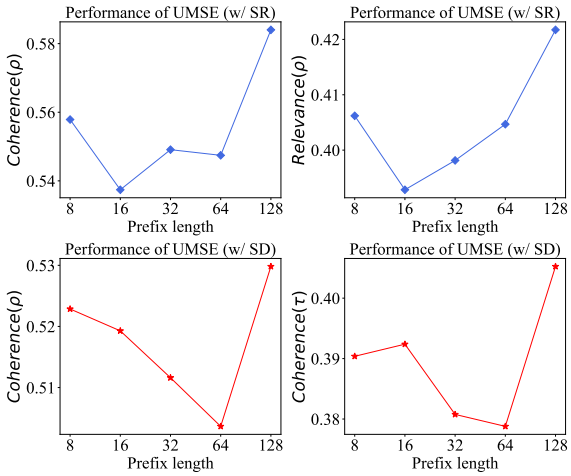

Figure 3: Performance across different prefix lengths.

users from having to use multiple models.

As illustrated in the related work § 2, some evaluators (*e.g.,* `UniEval` and `BARTScore`) focus on evaluating the summary in multi-dimension which model the specific dimension features and output *multiple scores*. Different from these methods, we focus on an orthogonal aspect that uses a unified model in multiple scenarios, and we only use *one score* to represent the summary quality. Thus, directly comparing with these multi-dimensional metrics is not fair. Since our unified multi-scenario evaluator is orthogonal to these multi-dimension evaluators, we will combine the multi-dimensional method into UMSE in future work.

Similar to our UMSE, `UniTE` is also a multi-scenario unified evaluation method for machine translation. However, `UniTE` achieves worse performance than UMSE, which demonstrates our assumption that the matching framework and the data construction method in `UniTE` are mainly focusing on the characteristic of translation. And we cannot simply use `UniTE` in the summarization task.

From the results of UMSE(Fusion) (w/ SDR) and UMSE (w/ SDR), we can find that the fusion model achieves better performance, and we will use the fusion method in our release version of UMSE. An extensive analysis of why the fusion method works better than directly concatenating Sum-Doc-Ref in the input of PLM is shown in the following section.

4.6 Discussions

Ablation Studies To verify the effectiveness of our proposed perturbed prefix tuning method, we employ an ablation model UMSE-PT in three scenarios. In this model, we mix the training datasets we constructed and jointly fine-tune one PLM for *all* scenarios. From the results shown in Table 1, we can find that UMSE-PT underperforms with the UMSE in all scenarios. Although using a shared pre-train language model can also transfer knowledge among these scenarios, these ablation studies demonstrate that using the shared continuous prefix tokens provides an explicit way to share common matching knowledge and it can boost the performance of the UMSE.

Moreover, we employ an intuitive experiment that separately fine-tunes a PLM for *each* scenario, and the results are shown in Table 3. Although the performance of the Sum-Ref drops slightly in terms of two dimensions, our proposed UMSE boosts the performance in the Sum-Doc scenario significantly. And boosting the performance of the Sum-Doc scenario is more valuable since evaluation in this scenario does not require any human annotating.

Analysis of Sum-Doc-Ref Fusion In § 3.4, we propose a variant model for the Sum-Doc-Ref scenario which directly fuses the scores of Sum-Doc and Sum-Ref to produce the score for the Sum-Doc-Ref scenario. In this section, we conduct experiments to explore which fusion method will lead to better performance. We employ four different fusion methods: (1) max method takes the maximum of s_{SD} and s_{SR} as s_{SDR}; (2) min method takes the minimum of s_{SD} and s_{SR}; (3) geometric mean fusion uses $\sqrt{s_{SD}s_{SR}}$ as s_{SDR}; and (4) arithmetic mean fusion employs $\frac{(s_{SD}+s_{SR})}{2}$. From Table 2, we can find that the arithmetic mean achieves the best performance, and we finally use the arithmetic mean fusion in the UMSE(Fusion).

Analysis of Perturbed Prefix Length To verify the effectiveness of our proposed perturbed prefix, we conduct experiments using the different lengths

of the prefix. From Figure 3, we can find that the performance of our UMSE gradually improved with the growth of the prefix length.

Analysis of Hallucination Detection To analyze the effectiveness of our model in detecting hallucinations, we conducted experiments on the dataset released by Maynez et al. (2020) and the results are shown in Table 4. According to the Spearman correlations on both faithful and factual, UMSE outperforms baselines, such as ROUGE, BERTScore, and QA, which demonstrates the ability of our proposed model in detecting hallucinations.

5 Conclusion

In this paper, we propose **U**nified **M**ulti-scenario **S**ummarization **E**valuation Model (UMSE) which is a unified multi-scenario summarization evaluation framework. UMSE can perform the semantic evaluation on three typical evaluation scenarios: (1) Sum-Ref; (2) Sum-Doc and (3) Sum-Doc-Ref using only one unified model. Since these scenarios have different input formats, we propose a perturbed prefix-tuning method that unifies these different scenarios in one model and it can also transfer knowledge between these scenarios. To train the UMSE in a self-supervised manner, we propose two training data construction methods without using any human annotation. Extensive experiments conducted on the benchmark dataset SummEval verify that the UMSE can achieve comparable performance with existing baselines.

Limitations

In this paper, we propose the evaluation model UMSE which can be used to evaluate the summary quality in three typical scenarios. However, in the summarization task, different annotators have different writing styles, and there might exist more than one good summary for one document. Moreover, there can be summaries that concentrate on different aspects of a document (*e.g.,* describing the location and room of a hotel). In the future, we aim to incorporate more scenarios (*e.g.,* multi-references and multi-aspects) into our unified evaluation method.

Ethics Statement

In this section, we would like to discuss the ethical concerns of our work. Our proposed method UMSE is a unified model for multi-scenario summarization evaluation and is designed to help humans efficiently evaluate summaries. And the sensitive information is masked while constructing the training data from CNN/DailyMail dataset.

Acknowledgements

We would like to express sincere thanks to the anonymous reviewers for their helpful comments. This research was supported by the Natural Science Foundation of China (T2293773, 62102234, 62272274, 62202271, 61902219, 61972234, 62072279), the National Key R&D Program of China with grant (No.2022YFC3303004, No.2020YFB1406704), the Key Scientific and Technological Innovation Program of Shandong Province (2019JZZY010129), the Tencent WeChat Rhino-Bird Focused Research Program (JR-WXG-2021411), the Fundamental Research Funds of Shandong University.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Wang Chen, Piji Li, and Irwin King. 2021. A training-free and reference-free summarization evaluation metric via centrality-weighted relevance and self-referenced redundancy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 404–414, Online. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, 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, Online. Association for Computational Linguistics.

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

Shen Gao, Xiuying Chen, Piji Li, Zhaochun Ren, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019a. Abstractive text summarization by incorporating reader comments. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6399–6406. AAAI Press.

Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Yan. 2020a. From standard summarization to new tasks and beyond: Summarization with manifold information. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4854–4860. ijcai.org.

Shen Gao, Zhaochun Ren, Yihong Eric Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019b. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 429–437. ACM.

Yang Gao, Wei Zhao, and Steffen Eger. 2020b. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *ArXiv preprint*, abs/1907.11692.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. 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 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for

ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

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

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

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.

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

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5602–5609. AAAI Press.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
7

☑ 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?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ 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.
Left blank.

C ☑ Did you run computational experiments?

3

☑ 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

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?
4

☑ 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?
5

☑ 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.)?
4

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

Left blank.

☐ 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.?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.