Union: An Unreferenced Metric for Evaluating Open-ended Story Generation

Jian Guan, Minlie Huang*
Department of Computer Science and Technology, Institute for Artificial Intelligence,
State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
j-guan19@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn

Abstract

Despite the success of existing referenced metrics (e.g., BLEU and MoverScore), they correlate poorly with human judgments for open-ended text generation including story or dialog generation because of the notorious one-to-many issue: there are many plausible outputs for the same input, which may differ substantially in literal or semantics from the limited number of given references. To alleviate this issue, we propose Union, a learnable ***UN**referenced metr**I**c for evaluating **O**pe**N**ded story generation*, which measures the quality of a generated story without any reference. Built on top of BERT, Union is trained to distinguish human-written stories from negative samples and recover the perturbation in negative stories. We propose an approach of constructing negative samples by mimicking the errors commonly observed in existing NLG models, including repeated plots, conflicting logic, and long-range incoherence. Experiments on two story datasets demonstrate that Union is a reliable measure for evaluating the quality of generated stories, which correlates better with human judgments and is more generalizable than existing state-of-the-art metrics.

Leading Context Jack was at the bar.
Reference By Human He noticed a phone on the floor. He was going to take it to lost and found. But it started ringing on the way. Jack answered it and returned it to the owner's friends.
Sample 1 (Reasonable, B=0.29, M=0.49, U=1.00) On the way out he noticed a phone on the floor. He asked around if anybody owned it. Eventually he gave it to the bartender. They put it into their lost and found box. **Sample 2 (Reasonable, B=0.14, M=0.27, U=1.00)** He had a drinking problem. He kept having more beers. After a while he passed out. When he waked up, he was surprised to find that he lost over a hundred dollars. **Sample 3 (Unreasonable, B=0.20, M=0.35, U=0.00)** He was going to get drunk and get drunk. The bartender told him it was already time to leave. Jack started drinking. Jack wound up returning but cops came on the way home.

Table 1: Generated story samples given the same leading context from ROCStories (Mostafazadeh et al., 2016). **B** stands for BLEU (Papineni et al., 2002), **M** for MoverScore (Zhao et al., 2019), and **U** for Union. A story can be reasonable even if it is dissimilar to the reference with a low BLEU score (B=0.14 in Sample 2), or unreasonable even if it has a large MoverScore (M=0.35 in Sample 3). In contrast, Union is more reliable for evaluating story generation.

1 Introduction

Significant advances have been witnessed with neural encoder-decoder paradigm (Sutskever et al., 2014), transformer-based architecture (Vaswani et al., 2017) and large-scale pretraining models (Devlin et al., 2019; Radford et al., 2019) in a wide array of natural language generation (NLG) tasks including machine translation (Bahdanau et al., 2015), story generation (Fan et al., 2018; Guan et al., 2020), and many more. However, the research is increasingly hindered by the lack of effective evaluation metrics, particularly for open-ended text generation tasks such as story generation.

Since human evaluation is time-consuming, expensive, and difficult to reproduce, the community commonly uses automatic metrics for evaluation. Previous studies in conditional language generation tasks (e.g., machine translation) have developed several successful referenced metrics, which roughly quantify the lexical overlap (e.g., BLEU (Papineni et al., 2002)) or semantic entailment (e.g., MoverScore (Zhao et al., 2019)) between a generated sample and the reference. However, such referenced metrics correlate poorly with

*Corresponding author

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 9157–9166,
November 16–20, 2020.

human judgments when evaluating open-ended text generation (Liu et al., 2016) due to the one-to-many nature (Zhao et al., 2017), as illustrated in Table 1. Specifically, a generated sample can be reasonable if it is coherent to the given input, and self-consistent within its own context but not necessarily being similar to the reference in literal or semantics, as shown in Sample 2 and 3.

To address the one-to-many issue, unreferenced metrics are proposed to measure the quality of a generated sample without any reference. Kannan and Vinyals (2017) presented a learnable, unreferenced metric which measures the text quality by learning to distinguish human-written texts from generated samples. However, the discriminator-based metric can easily lead to over-fitting to specific data (Garbacea et al., 2019) or model bias since the quality of generated texts varies substantially across different NLG models. As a matter of fact, the *generalization or robustness* issue is critical for any learnable metrics.

Therefore, we propose UNION, a learnable ***UN**referenced metr**I**c for evaluating **O**pen-e**N**ded story generation*. UNION learns to distinguish human-written stories from negative samples auto-constructed by generating perturbations of human-written stories. It is trained without dependence on specific NLG models or any human annotation, making it more generalizable to distribution drift (Sellam et al., 2020) than the discriminator-based metric and those metrics which learn from human preference (e.g., Adem (Lowe et al., 2017)). To capture commonly observed issues in generated stories, such as repeated plots, conflicting logic, and inter-sentence incoherence, we adopt four negative sampling techniques to construct negative samples, including repetition, substitution, reordering, and negation alteration. In addition, we design an auxiliary reconstruction objective for UNION, which recovers the perturbation from a negative sample. This objective is shown to further improve the performance of UNION.

Our contributions are summarized as follows:
I. We propose a learnable unreferenced metric UNION for evaluating open-ended story generation to alleviate the one-to-many issue of referenced metrics. UNION does not depend on any output of NLG models or human annotation.
II. Extensive experiments[1] show that UNION correlates better with human judgments than state-of-the-art metrics, and is more generalizable to data drift (samples from different datasets) and quality drift (samples with different quality levels).

2 Related Work

Automatic evaluation is crucial for language generation tasks. We roughly divide existing metrics into referenced, unreferenced, and hybrid metrics, according to whether they rely on human-written references when calculating the metric score.

Referenced metrics usually measure how similar a generated text is to the reference text. Therefore, they are developed mainly for conditional language generation tasks such as machine translation and text summarization, where plausible outputs are largely limited within the semantics of input. Commonly used referenced metrics include word-overlap based (e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004)) and embedding based metrics (e.g., BertScore (Zhang* et al., 2020), MoverScore (Zhao et al., 2019)). However, referenced metrics are reported to correlate poorly with human judgments in open-ended generation tasks including open-domain dialog generation (Liu et al., 2016) and story generation, where the input contains only limited information for generation, and there are many plausible outputs for the same input, which can vary substantially in literal or semantics.

Unreferenced metrics measure the quality of a sample without any reference. The most classic unreferenced metric is *perplexity*, which measures how likely a sample is generated by a given language model trained on human-written texts. However, recent work has shown that natural language is rarely the most probable text (Holtzman et al., 2020), and perplexity is inadequate to measure quality (Hashimoto et al., 2019). Therefore, perplexity may not indicate the actual text quality well. Discriminator-based metric (Kannan and Vinyals, 2017) measures how easily a discriminator distinguishes the generated samples from human-written texts. However, training such a discriminator can be easily over-fitted to a specific dataset, thereby leading to poor generalization and low correlation with human judgments (Garbacea et al., 2019). In addition to the above point-wise metrics which score an individual sample, Semeniuta et al. (2019) proposed the Fréchet InferSent Distance (FID) to evaluate the model-level quality and diversity of generated samples, by computing the Fréchet dis-

[1] All the codes and data are available at https://github.com/thu-coai/UNION.

tance between the Gaussian distribution fitted to human text embeddings and that to generated sample embeddings. However, in real data, the distribution of embeddings may be far from Gaussian. Recently, Zhou and Xu (2020) proposed to evaluate sample-level quality by comparing a pair of samples, and further adopted a skill rating system to evaluate model-level quality based on the sample-level pair-wise comparison. However, it is unlikely to evaluate a single sample without access to its references.

Hybrid metrics combine referenced and unreferenced metrics. For open-domain dialog system evaluation, Lowe et al. (2017) proposed a learnable metric Adem to learn from the human-annotated score of a response given its post and ground truth. However, such a metric shows very poor generalization and is not robust to easy attacks such as simple word substitution or random word shuffle (Sai et al., 2019). Furthermore, RUBER and its variants (Tao et al., 2018; Ghazarian et al., 2019) evaluate a response by directly averaging a non-learnable referenced embedding similarity score and a learnable unreferenced post-response relatedness score that is learned by applying negative sampling without human annotations. However, merely measuring input-output relatedness is not sufficient for evaluating long text generation, as the intrinsic coherence and consistency within the generated text is a critical factor. Additionally, some metrics which learn from human preference achieve substantial results in conditional language generation, e.g., RUSE (Shimanaka et al., 2018) and BLEURT (Sellam et al., 2020). RUSE trained a regression model to score a reference-candidate pair using their sentence embeddings. And BLEURT used multiple automatic metrics (e.g., BLEU) as supervision signals for pretraining on synthetic data, and was fine-tuned on human judgments. However, BLEURT heavily relies on the quality of automatic metrics, but there are yet no such reliable metrics for open-ended text generation.

3 Methodology

UNION is expected to measure the overall quality of a generated story. In this section, we begin with common issues that can be observed in the output of NLG models. We then propose four negative sampling techniques based on the observations. Afterward, we introduce how UNION is trained and used for story evaluation. The overall paradigm of UNION is shown in Figure 1.

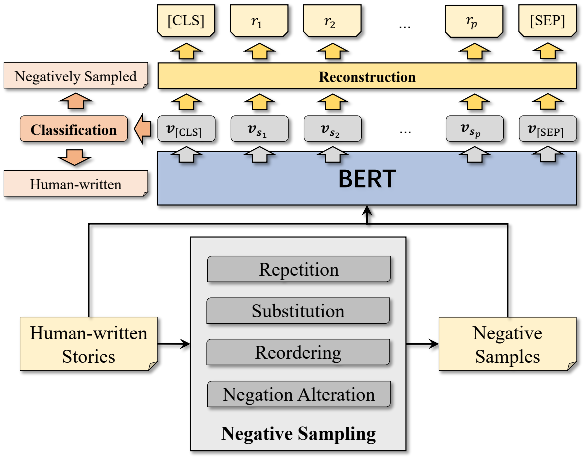

Figure 1: Overview of the UNION metric. UNION is trained to distinguish the human-written stories from the negative samples constructed by four negative sampling techniques, as well as to reconstruct the original human-written stories.

3.1 Empirical Observations

The key aspect of UNION is the construction of negative samples, which provides a range of lexical, syntactic, and semantic variations to simulate the errors made by NLG models. Therefore, we first present our empirical observations regarding the question *"What makes a story unreasonable for NLG models?"*.

We analyzed 381 unreasonable stories generated by various NLG models like Plan&Write (Yao et al., 2019) and fine-tuned GPT-2 (Radford et al., 2019) base on ROCStories (Mostafazadeh et al., 2016), and summarized four major types of errors, including **repeated plots** (repeating similar texts), **poor coherence** (with unrelated keywords or events but a reasonable main plot), **conflicting logic** (wrong causal or temporal relationship), and **chaotic scenes** (difficult to understand or with multiple previous errors). To facilitate understanding of the error types, we resorted to manual annotation of all the unreasonable stories. And seven annotators were hired for each story (see the full details in Section 4.2). In addition to the four error types, we also provide annotators with an option **Others**. We summarize the proportion of stories annotated with different error types in Table 2[2].

We can see that the four error types are the major issues of unreasonable stories, which provides

[2] Note that these human annotations are only used in test of UNION.

Type	**Repe**	**Cohe**	**Conf**	**Chao**	**Others**
Prop (%)	44.1	56.2	67.5	50.4	12.9

Table 2: Error type **Prop**ortions of 381 unreasonable stories, including **Repe**ated plots/poor **Cohe**rence/**Conf**licting logic/**Chao**tic scenes/**Others**.

rationales of constructing negative samples for evaluating generated stories. Besides, all the Spearman correlations between every two error types are less than 0.15 (p-value > 0.01), suggesting that different error types correlate weakly with each other. Furthermore, the stories annotated with 1/2/3/4 errors constitute 23.36%/36.48%/34.65%/4.46% of the annotated stories, respectively. Most of the unreasonable stories have more than one error, which motivates us to simultaneously apply multiple sampling techniques to construct negative samples.

3.2 Constructing Negative Samples

We construct negative samples to cover as many aforementioned issues of unreasonable stories as possible. Since using machine-generated texts as negative samples will easily lead to poor generalization (over-fitting to specific data or model bias (Garbacea et al., 2019)), we devise four negative sampling techniques to automatically construct a large number of negative samples from human-written stories as follows:

Repetition: Generating repetitive texts is commonly observed in many state-of-the-art NLG models (Fan et al., 2018; Radford et al., 2019), where the models focus repeatedly on what they have recently generated, particularly with maximum-likelihood based decoding strategies (Holtzman et al., 2020). To address the issue, we introduce lexical and sentence-level repetition to construct negative samples using two policies—we either repeat an N-gram (N=1,2,3,4) in a random sentence, or randomly select a sentence to repeat and remove the following sentence to keep the sentence number unchanged.

Substitution: The coherence of a story is mainly embodied through the relationship between keywords in the context (Clark et al., 2018; Guan et al., 2020). Therefore, we create incoherent samples by random keywords and sentence substitution, respectively at word level and sentence level. For word-level substitution, we replace random 15% keywords in a story with their corresponding antonyms (e.g., replace “deny” with “confirm”), otherwise with another random keyword sampled from all the keywords of the same part-of-speech (POS), according to the mention frequency. We use the commonsense knowledge base ConceptNet (Speer and Havasi, 2012)[3] for keyword recognition and antonym query. ConceptNet consists of commonsense triples like `(h, r, t)`, meaning that the head concept `h` has a relation `r` with the tail concept `t`, e.g., `(evaluation, IsA, judgment)`. We regard those words which are heads or tails in ConceptNet as keywords. And given an keyword, we look up those keywords as its antonyms with which have negated relations, including `Antonym`, `NotDesires`, `NotCapableOf`, and `NotHasProperty`. If no antonym is found for a keyword, we perform replacement with a random keyword of the same POS. And we adopt NLTK[4] for POS tagging.

For sentence-level substitution, we randomly replace a sentence in a story with another one sampled from the rest of stories in the dataset.

Reordering: Conflicting logic usually results from wrong causal relationship and temporal dependency in the context. Therefore, we randomly reorder the sentences in a story to create negative stories with conflicting plot.

Negation Alteration: Negation words such as “not” are crucial for language generation tasks because they may flip the semantics of a sentence, which is also an important cause of conflicting logic. We perform negation alteration by adding or removing negation words using rules for different types of verbs[5].

Since there may be multiple error types in a generated story, we apply different sampling techniques simultaneously to construct a negative sample. We first sample the number (n) of techniques from {1,2,3,4} with a distribution {50%, 20%, 20%, 10%}. We then sample a technique without replacement from {repetition, substitution, reordering, negation alteration} with a distribution {10%, 30%, 40%, 20%} until the total number of techniques (n) is reached. Last, we apply the sampled techniques on a human-written story to obtain a perturbated sample. A constructed example is shown in Table 3.

[3] `http://www.conceptnet.io/`

[4] `http://nltk.org/`

[5] The details are shown in the supplementary material.

Leading Context Ken was out jogging one morning.
Reference By Human The weather was crisp and cool. Ken felt good and energetic. He decided to keep jogging longer than normal. Ken went several more miles out of his way.
Auto-Constructed Negative Sample The weather was crisp and cool *and cool.* Ken felt bad and energetic. **Ken DID NOT GO several more miles out of his way. He decided to keep jogging longer than normal**.

Table 3: An example of negative sample construction. The repeated bigram is in *italic*, the substituted keyword is underlined, the reordered sentences are indicated in **bold**, and the altered negation words are CAPITALIZED.

3.3 Modeling

Let $\{\boldsymbol{s}_n, \boldsymbol{r}_n, y_n\}_{n=1}^N$ denote the training dataset of size N for training the UNION metric, where $\boldsymbol{s}_n$ is a human-written story or an auto-constructed negative sample, $\boldsymbol{r}_n$ is the corresponding original story of $\boldsymbol{s}_n$. If $\boldsymbol{s}_n$ is a negative sample, $y_n = 0$, otherwise $y_n = 1$ where $\boldsymbol{s}_n$ is exactly the same as $\boldsymbol{r}_n$ in this case. $y_n \in \{0, 1\}$ indicates whether $\boldsymbol{s}_n$ is written by human. For better story understanding, we leverage BERT (Devlin et al., 2019) to obtain contextualized representations of the input. Given a story $\boldsymbol{s}_n = (s_1, s_2, \cdots, s_p)$ of length p (each s_i is a word), BERT outputs a sequence of contextualized vectors:

$$\boldsymbol{v}_{[\mathrm{CLS}]}, \boldsymbol{v}_{s_1}, \cdots, \boldsymbol{v}_{s_p}, \boldsymbol{v}_{[\mathrm{SEP}]} = \mathrm{BERT}(\boldsymbol{s}_n), \quad (1)$$

where $\boldsymbol{v}_{[\mathrm{CLS}]}$ and $\boldsymbol{v}_{[\mathrm{SEP}]}$ are the representation for the special tokens [CLS] and [SEP], respectively. We add a task-specific linear layer on top of the [CLS] vector to predict the UNION score, indicating the probability that $\boldsymbol{s}_n$ is written by human:

$$\hat{y}_n = \mathrm{sigmoid}(\mathbf{W}_c \boldsymbol{v}_{[\mathrm{CLS}]} + \mathbf{b}_c), \quad (2)$$

where $\mathbf{W}_c$ and $\mathbf{b}_c$ are trainable parameters. We use the cross entropy loss to optimize the prediction objective as follows:

$$\mathcal{L}_n^C = -y_n \log \hat{y}_n - (1 - y_n) \log (1 - \hat{y}_n). \quad (3)$$

In addition to the main prediction task, we devise an auxiliary reconstruction task which requires to reconstruct the corresponding human-written story $\boldsymbol{r}_n$ from perturbated story $\boldsymbol{s}_n$. Therefore, we add an additional linear layer at the last layer of BERT, which takes as input the vectors output from the last transformer block and computes a probability distribution over the entire vocabulary through a softmax layer, formally as follows:

$$P(\hat{r}_i | \boldsymbol{s}_n) = \mathrm{softmax}(\mathbf{W}_r \boldsymbol{v}_{s_i} + \mathbf{b}_r), \quad (4)$$

where $\hat{r}_i$ is the predicted i-th token, $\mathbf{W}_r$ and $\mathbf{b}_r$ are the parameters of the additional linear layer. Then the model is trained by minimizing the negative log-likelihood:

$$\mathcal{L}_n^R = -\frac{1}{p} \sum_{i=1}^{p} \log P(\hat{r}_i = r_i | \boldsymbol{s}_n), \quad (5)$$

where r_i is the i-th token in human-written story $\boldsymbol{r}_n$. The combined loss function $\mathcal{L}$ of the full model is computed as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} (\mathcal{L}_n^C + \lambda \mathcal{L}_n^R), \quad (6)$$

where λ is an adjustable hyperparameter.

We fine-tune all the parameters of UNION on the training dataset, including the BERT and the two additional linear layers. In practical use, UNION can measure the quality of a new generated sample $\hat{\boldsymbol{s}}$ by taking $\hat{\boldsymbol{s}}$ as input to predict the corresponding score $\hat{y}$.

4 Experiment

We conducted extensive experiments to evaluate UNION on two story datasets. First, we compared UNION against existing text generation metrics. Then, we assessed its generalization on distribution drifts, including dataset drift and quality drift. Last, we measured the effect of each negative sampling technique with ablation studies.

4.1 Baselines

We compared UNION with the following three kinds of metrics as baselines:

Referenced metrics: sentence ***BLEU*** score (geometric mean of 1-gram up to 4-gram) (Papineni et al., 2002) to measure the lexical similarity between a candidate sample and its reference, and ***MoverScore*** (Zhao et al., 2019) to measure the semantic similarity.

Unreferenced metrics: ***Perplexity***[6] computed by the GPT-2 model (Radford et al., 2019), and a discriminative evaluator (***DisScore***) (Kannan and

[6] We take the minus of perplexity for all the following experiments to ensure a higher value means better quality.

Vinyals, 2017) that is trained based on BERT to distinguish generated samples from human-written stories.

Hybrid metrics: ***RUBER-BERT*** (Ghazarian et al., 2019) which improves the original RUBER (Tao et al., 2018) with contextualized embeddings from BERT, and the supervised metric ***BLEURT*** (Sellam et al., 2020) that is fine-tuned on human judgments after pretraining on large-scale synthetic data with multiple automatic metrics as supervision signals.

In addition, we also reported the performance of the referenced and unreferenced versions in RUBER-BERT, denoted as ***RUBER$_r$-BERT*** and ***RUBER$_u$-BERT***, respectively.

We set the parameters of UNION by following the uncased base version of Devlin et al. (2019): the transformer has 12 layers, 768 dimensional hidden states, and 12 attention heads. We used batch size 10, and learning rate 5e-5. The scale factor λ is set to 0.1. We directly used public pretrained parameters of BERT[7] or GPT-2[8] (base version) for all the baselines.

4.2 Data Preparation

We used two datasets for evaluation, ROCStories (**ROC** for short) (Mostafazadeh et al., 2016) and WritingPrompts (**WP**) (Fan et al., 2018). The ROC dataset contains 98,161 five-sentence human-written stories, with an average length of 49.4 words. To achieve better generalization performance, we followed Guan et al. (2020) to make delexilization by masking all the male/female/unknown names with placeholders [MALE]/[FEMALE]/[NEUTRAL], respectively.

The WP dataset consists of 303,358 stories paired with writing prompts collected from an online forum. The average length of the prompt/story is 28.4/734.5 respectively, much longer than those in ROC. Since it is still challenging for state-of-the-art NLG models to maintain a reasonable plot through the whole story, and hard to obtain acceptable annotation agreement in manual evaluation of long stories, we retained about 200 words (with correct sentence boundary) from the start and truncated the rest in WP for subsequent experiments.

We randomly selected 90%/5%/5% stories from both datasets for training/validation/test of UNION and learnable baseline metrics, and created the evaluation set for all the metrics by generating stories based on the test sets of the datasets with state-of-the-art story generation models. The story generation models include fusion convolutional seq2seq model (Fan et al., 2018), plan&write (Yao et al., 2019), fine-tuned GPT-2 (Radford et al., 2019), and knowledge-enhanced GPT-2 (Guan et al., 2020).

The data statistics are shown in Table 4. The number of negative samples for learning the metrics when necessary is the same as that of human-written stories on each dataset. Specifically, we created negative samples for DisScore by generating stories with above NLG models. For RUBER$_u$-BERT, a given leading context is appended by a randomly sampled continuation. All the stories in the evaluation set are manually labeled. In addition, we annotated another 400 stories in ROC and 200 in WP for training BLEURT[9]. Seven annotators were hired to judge the quality of each story with a binary score (1 for a reasonable story, and 0 otherwise). Furthermore, we asked annotators to label the error type of a story if it is labeled as unreasonable, including repeated plots, poor coherence, conflicting logic, chaotic scenes, and others. We resorted to Amazon Mechanical Turk (AMT) for annotation, and the average score of the seven annotators is treated as the final score. We provide the full details of the instruction for annotators in the supplementary file.

Split	Metrics	ROC	WP	NS
Train/ Validate	Perplexity DisScore RUBER$_u$ UNION	88,344/ 4,908	272,600/ 15,620	✗ ✓ ✓ ✓
	BLEURT	360†/40†	180†/20†	✗
Test	All metrics	400†	200†	N/A

Table 4: Data statistics. **RUBER$_u$** is short for **RUBER$_u$-BERT**. **NS** (Negative Sampling) means whether a metric requires negative samples for training/validation. † means the stories are generated by NLG models and manually annotated.

4.3 Correlation Results

Correlation analysis has been widely used to evaluate automatic metrics for language generation (Tao et al., 2018; Sellam et al., 2020). We employed UNION and other metrics to score the collected samples, and then calculated the Pearson (r),

[7] https://github.com/google-research/bert

[8] https://github.com/openai/gpt-2

[9] BLEURT is first initialized with the pretrained parameters (https://github.com/google-research/bleurt) and then fine-tuned on our annotated stories.

Metrics		ROC			WP		
		r	ρ	τ	r	ρ	τ
Referenced	**BLEU**	0.0299	0.0320	0.0231	0.1213	0.0941	0.0704
	MoverScore	0.1538*	0.1535*	0.1093*	0.1613	0.1450	0.1031
	RUBER$_r$-BERT	0.0448	0.0517	0.0380	0.1502	0.1357	0.0986
Unreferenced	**Perplexity**	0.2464*	0.2295*	0.1650*	-0.0705	-0.0479	-0.0345
	RUBER$_u$-BERT	0.1477*	0.1434*	0.1018*	0.1613	0.1605	0.1157
	DisScore	0.0406	0.0633	0.0456	0.0627	-0.0234	-0.0180
	UNION	**0.3687***	**0.4599***	**0.3386***	**0.3663***	**0.4493***	**0.3293***
	-Recon	0.3101*	0.4027*	0.2927*	0.3292*	0.3786*	0.2836*
Hybrid	**RUBER-BERT**	0.1412*	0.1395*	0.1015*	0.1676	0.1664	0.1194
	BLEURT	0.2310*	0.2353*	0.1679*	0.2229*	0.1602	0.1180

Table 5: Correlation with human judgments on ROC and WP datasets. $r/\rho/\tau$ indicates the Pearson/Spearman/Kendall correlation, respectively. The best performance is highlighted in **bold**. The correlation scores marked with * indicate the result significantly correlates with human judgments (p-value<0.01).

Spearman (ρ) and Kendall (τ) correlation coefficients between model evaluation and human judgments. Pearson's r estimates linear correlation while Spearman's ρ and Kendall's τ estimate monotonic correlation, and τ is usually more insensitive to abnormal values than ρ. We used the standard statistical package `stats` in `SciPy`[10] for correlation calculation and significance test.

As summarized in Table 5, the referenced metrics correlate worse with human judgments, particularly for BLEU which is based on lexical similarity. Measuring the semantic similarity instead (MoverScore, RUBER$_r$-BERT) can improve the correlation but is still limited, indicating that referenced metrics are not competitive for evaluating open-ended language generation. Perplexity is ineffective on WP because the generated stories in the dataset are much longer and hence suffer from more serious repetition errors than those in ROC, which easily results in low perplexity (i.e., high minus perplexity) (Holtzman et al., 2020) but poor human judgment scores. Furthermore, UNION outperforms other baselines including the supervised metric BLEURT by a large margin, which also demonstrates the advantage of unreferenced metrics. Besides, removing the reconstruction training objective (-Recon) leads to remarkably worse correlation, indicating that the auxiliary task further improves the performance of UNION.

4.4 Generalization to Dataset and Quality Drift

It is extremely important for learnable metrics to deal with dataset drift and quality drift (Sellam et al., 2020). Specifically, a generalizable metric is expected to reliably evaluate outputs from different datasets even without re-training. Moreover, since the quality of generated samples can vary significantly across NLG models, a reliable metric should be able to evaluate samples of different quality levels. Therefore, we conducted experiments to assess the generalization ability of UNION in this section.

Metrics	r	ρ	τ
Training: WP **Test**: ROC			
Perplexity	-0.0015	0.0149	0.0101
RUBER$_u$-BERT	-0.0099	-0.0162	-0.0110
BLEURT	0.1326*	0.1137*	0.0828*
UNION	**0.1986***	**0.2501***	**0.1755***
-Recon	0.1704*	0.2158*	0.1523*
Training: ROC **Test**: WP			
Perplexity	0.0366	0.0198	0.0150
RUBER$_u$-BERT	0.1392	0.1276	0.0912
BLEURT	0.1560	0.1305	0.0941
UNION	**0.2872***	**0.2935***	**0.2142***
-Recon	0.2397*	0.2712*	0.1971*

Table 6: Correlation results in the dataset drift setting where the metrics are trained on one dataset and then used for the other one.

To assess the generalization to dataset drift, we first trained the learnable metrics on ROC and then directly used them to evaluate generated stories from WP, and vise versa. Table 6 shows the Pearson correlation with human judgments in this setting. Compared with the results in Table 5, all the metrics trained on one dataset have remarkable drops in correlation when they are used for the other dataset because the two datasets are significantly different in length and topic. Nevertheless, UNION performs more robustly than other metrics, with much bet-

[10] https://docs.scipy.org/doc/scipy/reference/stats.html

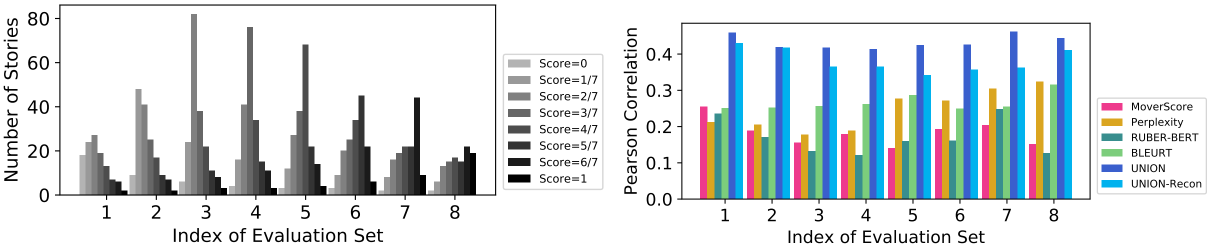

Figure 2: Generalization over different biased test sets. Left: distribution of stories of different annotation scores in different test sets. Right: the Pearson correlation of different metrics with human judgments on different test sets, where UNION-Recon denotes UNION without the reconstruction task.

Evaluation Set	**All Samples (400)**	**Reasonable Samples (19) + Unreasonable Samples with**			
		Repe (24)	**Cohe (38)**	**Conf (61)**	**Chao (23)**
UNION	0.3687	0.6943	0.5144	0.4571	0.6744
-Repetition	0.3167 (↓**14%**)	0.4743 (↓**32%**)	0.5308 (↑3%)	0.4316 (↓6%)	0.6561 (↓3%)
-Substitution	0.3118 (↓**15%**)	0.7034 (↑1%)	0.4185 (↓**19%**)	0.4468 (↓2%)	0.5850 (↓**13%**)
-Reordering	0.2302 (↓**38%**)	0.6546 (↓6%)	0.5077 (↓1%)	0.3507 (↓**23%**)	0.5393 (↓**20%**)
-Negation Alteration	0.3304 (↓**10%**)	0.6665 (↓4%)	0.4987 (↓3%)	0.3946 (↓**14%**)	0.5176 (↓**23%**)

Table 7: Pearson correlation with different negative sampling techniques. The numbers in parentheses denote the number of stories. The error types include **Repe**ated plots, poor **Cohe**rence, **Conf**licting logic, and **Chao**tic scenes. The proportions in parentheses indicate the relative change with respect to UNION (the first row).

ter correlation with human judgments. Moreover, our method of constructing negative examples is generalizable to the two datasets.

To assess the generalization of UNION to quality drift, we created biased test sets from ROC by sampling stories of different quality levels with different probabilities. Specifically, the annotation score of each story ranges from 0 to 1 (i.e., $0, \frac{1}{7}, \frac{2}{7}, \cdots, 1$) since there are seven annotators for each sample. We then created 8 biased sets, indexed from 1 to 8 with variable I. For the I^{th} set, we sampled the stories whose annotation score is $\frac{k}{7}$ with a probability of $\frac{1}{|I-k|+1}$ where $k \in \{0, 1, \cdots, 7\}$. In this way, the 8 sets have different distributions of stories with different qualities[11], as shown in Figure 2 (left).

We then computed the Pearson correlation of different metrics with human judgments on the 8 sets. Results in Figure 2 (right) show that: **I.** UNION has higher correlation than other metrics on all the biased sets. **II.** UNION is more reliable and robust than other metrics, with much less variance. For instance, MoverScore performs much better on Set #1 (with more low-quality stories) than on Set #8 (with more high-quality stories). Interestingly, Perplexity performs much better on high-quality sets than on low-quality ones, because high-quality stories are closer to human-written stories from which a language model learns. **III.** The ablated UNION without the reconstruction objective has lower correlation and larger variance, indicating that the auxiliary task can improve the discriminative and generalization ability.

4.5 Ablation Studies

To understand the effect of each negative sampling technique, we conducted ablation tests on ROC dataset. Each time we ablated one technique of constructing negative samples, re-trained UNION on the constructed data, and evaluated it on five evaluation sets: all 400 samples, and four other sets where each contains 19 reasonable samples and other unreasonable samples of some error type. The error type of a story is decided if at least three of seven annotators annotate the same error type.

Table 7 shows the Pearson correlation results. UNION is remarkably better than its ablated version on the all-sample set, indicating the necessity of the four techniques for constructing negative samples. Reordering seems to be the most important technique, which agrees with our observation that conflicting logic is the major issue in existing story generation models. Furthermore, as expected, the correlation drops remarkably on the evaluation set of some error type if without the corresponding negative sampling technique. Interestingly, it is easier for UNION to evaluate repetitive/chaotic stories, which seem to be easier cases in story generation.

[11] We assume that the annotation score $\frac{k}{7}$ approximates the quality level.

5 Conclusion

We present UNION, an unreferenced metric for evaluating open-ended story generation. UNION is trained to distinguish human-written stories from auto-constructed negative samples and to recover the perturbation in negative samples. Extensive experiments show that UNION outperforms state-of-the-art metrics in terms of correlation with human judgments on two story datasets, and is more robust to dataset drift and quality drift. Results also show the effectiveness of the proposed four negative sampling techniques. As future work, we will explore the similar idea of designing unreferenced metrics for dialog generation.

Acknowledgments

This work was jointly supported by the NSFC projects (Key project with No. 61936010 and regular project with No. 61876096), and the Guoqiang Institute of Tsinghua Universitywith Grant No. 2019GQG1. We thank THUNUS NExT Joint-Lab for the support.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Elizabeth Clark, Yangfeng Ji, and Noah A. Smith. 2018. Neural text generation in stories using entity representations as context. In *NAACL*, pages 1631–1640.

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.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Cristina Garbacea, Samuel Carton, Shiyan Yan, and Qiaozhu Mei. 2019. Judge the judges: A large-scale evaluation study of neural language models for online review generation. 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 3959–3972.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. 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 1689–1701.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Anjuli Kannan and Oriol Vinyals. 2017. Adversarial evaluation of dialogue models. *arXiv preprint arXiv:1701.08198*.

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.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6220–6227.

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

Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. 2019. On accurate evaluation of GANs for language generation.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

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

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.

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.

Wangchunshu Zhou and Ke Xu. 2020. Learning to compare for better training and evaluation of open domain natural language generation models. In *AAAI*, pages 9717–9724.