Dialogue Rewriting via Skeleton-Guided Generation

Chunlei Xin[1,3], Hongyu Lin[1,*], Shan Wu[1,3], Xianpei Han[1,2], Bo Chen[1,5,6], Wen Dai[4], Shuai Chen[4], Bin Wang[4], Le Sun[1,2,*]

[1]Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China
[2]State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Xiaomi AI Lab, Xiaomi Inc., Beijing, China
[5]School of Information Engineering, Minzu University of China, Beijing, China
[6]National Language Resources Monitoring and Research Center for Minority Languages, Beijing, China
{chunlei2021, hongyu, wushan2018, xianpei, chenbo, sunle}@iscas.ac.cn
{daiwen, chenshuai3, wangbin11}@xiaomi.com

Abstract

Dialogue rewriting aims to transform multi-turn, context-dependent dialogues into well-formed, context-independent text for most NLP systems. Previous dialogue rewriting benchmarks and systems assume a fluent and informative utterance to rewrite. Unfortunately, dialogue utterances from real-world systems are frequently noisy and with various kinds of errors that can make them almost uninformative. In this paper, we first present Real-world Dialogue Rewriting Corpus (RealDia), a new benchmark to evaluate how well current dialogue rewriting systems can deal with real-world noisy and uninformative dialogue utterances. RealDia contains annotated multi-turn dialogues from real scenes with ASR errors, spelling errors, redundancies and other noises that are ignored by previous dialogue rewriting benchmarks. We show that previous dialogue rewriting approaches are neither effective nor data-efficient to resolve RealDia. Then this paper presents Skeleton-Guided Rewriter (SGR), which can resolve the task of dialogue rewriting via a skeleton-guided generation paradigm. Experiments show that RealDia is a much more challenging benchmark for real-world dialogue rewriting, and SGR can effectively resolve the task and outperform previous approaches by a large margin.

Introduction

Dialogue is the primary mechanism for human interaction, which is multi-turn, context-dependent, and frequently informal (Pangaro and Dubberly 2014). People tend to produce brief, fragmented utterances in dialogues rather than longer, completed sentences in normal documents (Carbonell 1983). Therefore, many utterances in dialogues can only be understood when they are put in the entire dialogue context. Unfortunately, most NLP systems are designed for well-formed, context-independent texts, which makes them inappropriate to deal with informal and context-dependent dialogues. To narrow the gap, current studies mostly focus on developing ***dialogue-specialized paradigm***, which expands specific downstream tasks from single-sentence inputs to multi-turn dialogue inputs, and designs complicated

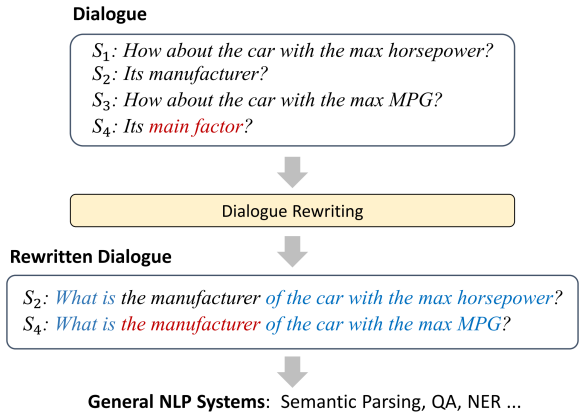

Figure 1: An example shows how dialogue rewriting handles multi-turn dialogues. Key information restored from the dialogue is marked in blue, and the automatic speech recognition error and its correction are in red.

mechanisms which can take richer context information into consideration (Wu et al. 2017; Zhang et al. 2019; Wang et al. 2021). However, dialogue-specialized models are usually hard to design, and it is time-consuming and unacceptably costly to construct dialogue-specialized models for various downstream tasks. Therefore, how to more effectively deal with noisy, heavily context-dependent utterances in dialogues is a critical challenge for dialogue understanding.

Recently, the task of *dialogue rewriting* was proposed to resolve the above-mentioned challenge. Dialogue rewriting transforms inter-dependent and informal utterances in multi-turn dialogues into well-formed, context-independent and semantically completed sentences, and therefore dialogue utterances can be seamlessly restructured into well-formed inputs for current NLP systems. For example, given the utterance "Its manufacturer?" of utterance S_2 in Figure 1, dialogue rewriting models need to distill relevant information "the car" and "the max horsepower", and output the sentence "What is the manufacturer of the car with the max horsepower?", which can accurately express the speaker intent of the original utterance under the dialogue context.

* Corresponding authors.

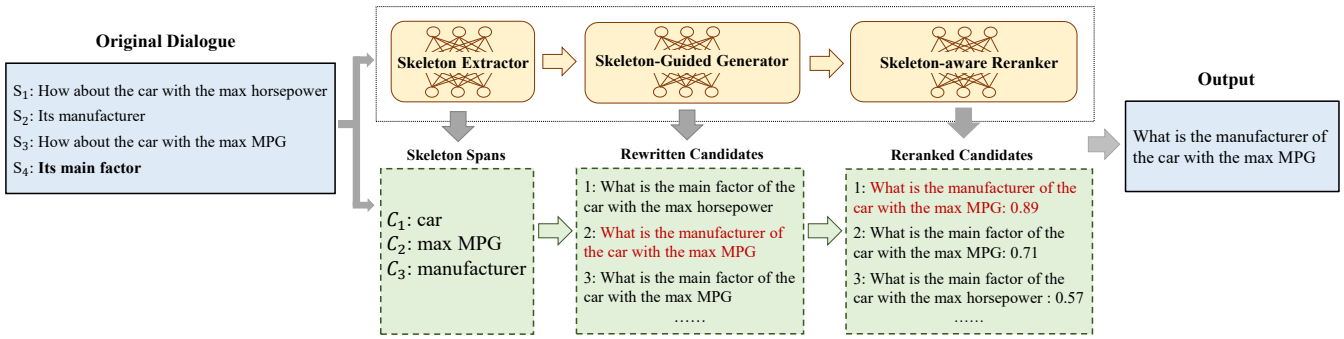

Figure 2: Overall architecture of the proposed Skeleton-Guided Rewriter. The sentence that needs to be rewritten is shown in bold and the golden is marked in red.

Previous dialogue rewriting benchmarks assume a fluent and informative utterance to rewrite. However, dialogue utterances from real-world systems are frequently informal, noisy and with various kinds of errors that can make them uninformative. For example, many dialogues transcribed from speeches contain automatic speech recognition (ASR) errors, and dialogues from user textual inputs frequently contain spelling errors. For the instance in Figure 1, the ASR error "main factor" can make this utterance incomprehensible and uninformative, and it is impossible to directly obtain the correct user intent by only rewriting this utterance. As a result, as shown in 1, previous benchmarks either filter out dialogues with errors and only consider fluent and informative dialogues (Su et al. 2019; Elgohary, Peskov, and Boyd-Graber 2019; Quan et al. 2019), or ignore the errors and allow the output to still contain these errors (Pan et al. 2019). Unfortunately, the former choice does not match the nature of real-world dialogue rewriting, while the latter fails to meet the requirements of downstream NLP systems. Consequently, current dialogue rewriting benchmarks and systems can not fully meet the requirements of dialogue rewriting in real-world applications.

In this paper, we present Real-world Dialogue Rewriting Corpus (RealDia), a new dialogue rewriting benchmark collected from real-world dialogues. Compared with previous dialogue rewriting benchmarks, RealDia annotates open-domain, multi-turn dialogues from real scenes with ASR errors, spelling errors, redundancies and other noises. Dialogue rewriting models need not only to deal with ellipsis and co-reference as previous dialogue rewriting benchmarks, but also to infer the underlying semantics of uninformative utterances based on the dialogue contexts. Furthermore, model outputs should be natural language expressions with high fluency, high coverage and high consistency that can meet the input standards for various downstream NLP systems. Such requirements pose remarkable challenges to previous dialogue rewriting systems, as they are designed based on the assumption of rewriting fluent and informative utterances. As a result, previous approaches can not achieve desirable performances on real-world dialogues in RealDia.

To this end, we further propose ***Skeleton-Guided Rewriter (SGR)***, which can effectively and efficiently resolve real-world dialogue rewriting using a skeleton-guided generation framework. The main idea behind SGR is to ask the model to pay more attention to critical information (i.e., the skeleton) across the entire dialogue session, which can reduce the dependence on the utterance itself to resolve the uninformative utterance challenge. Figure 2 shows the overall architecture of SGR, which contains three critical steps: 1) **Dialogue Skeleton Extraction**, which identifies the key information that needs to be covered in the rewritten sentence; 2) **Skeleton-Guided Generation**, which generates fluent candidate sentences under the guidance of skeleton; 3) **Skeleton-aware Reranking**, which selects the best candidate by measuring the fluency, coverage and semantic consistency with the original dialogue of each candidate.

We evaluate the effectiveness of SGR on both RealDia and previous dialogue rewriting benchmarks. Experiments show that previous approaches can not effectively solve the task of real-world dialogue rewriting on RealDia because of their inherent informative utterance assumption. And by leveraging the skeleton across the entire dialogue, SGR can simultaneously resolve ellipsis, co-reference, redundancies and various kinds of errors, and therefore achieve state-of-the-art performance on both RealDia and previous dialogue rewriting benchmarks.

Generally speaking, the main contributions of this paper can be summarized as:

- We construct RealDia, a new dialogue rewriting benchmark collected from real-world dialogues to evaluate how well current systems can rewrite real-world noisy and uninformative dialogue utterances. To the best of our knowledge, RealDia is the first dialogue rewriting benchmark that considers uninformative utterances and various kinds of errors in real-world dialogues.
- We design Skeleton-Guided Rewriter (SGR), a skeleton-guided generation framework that can effectively and efficiently rewrite uninformative utterances in real-world dialogues. To the best of our knowledge, SGR is the first work that attempts to leverage skeleton-guided generation framework to better resolve uninformative utterance challenge in dialogue rewriting.
- Experiments show that RealDia is much more challenging than previous dialogue rewriting benchmarks, and

Dataset	Fluent Utterances	Informative Utterances	Legal Sentence as Result
TASK	Yes	Yes	Yes
CANARD	Yes	Yes	Yes
REWRITE	Yes	Yes	Yes
Restoration	No Limit	No Limit	No Limit
RealDia	No Limit	No Limit	Yes

Table 1: Comparison between RealDia and existing dialogue rewriting benchmarks. Previous benchmarks either only collected fluency and informative dialogues, or ignored noise and allowed outputs to remain illegal.

SGR achieves state-of-the-art performance on both RealDia and previous benchmarks. This demonstrates the necessity of RealDia and the effectiveness of SGR.

Related Work

Dialogue rewriting aims to transform inter-dependent and informal utterances in multi-turn dialogues into well-formed, context-independent and semantically completed sentences. Along this line, Su et al. (2019) collect a rewriting dataset for co-reference resolution and information completion in multi-turn dialogues and propose a pointer-based rewriter. Pan et al. (2019) collect a Restoration-200K dataset and propose a cascaded pick-and-combine model. Quan et al. (2019) construct a dataset with both ellipsis and co-reference annotation and propose an end-to-end generative resolution model. Liu et al. (2020) formulate incomplete utterance rewriting as a semantic segmentation task. Mele et al. (2021) propose adaptive utterance rewriting strategies for better conversational information retrieval. Hao et al. (2021) propose a tagging-based approach that predicts edit actions to rewrite incomplete utterances, based on which Jin et al. (2022) propose a hierarchical context tagger to expand the coverage and shrink search space. Recently, Si, Zeng, and Chang (2022) propose the query-enhanced network, which consists of a query template construction module and an edit operation scoring network. Inoue et al. (2022) jointly optimize picking important tokens and generating rewritten utterances.

Although dialogue rewriting has recently raised wide attention, previous dialogue rewriting benchmarks assume a fluent and informative utterance to rewrite, and therefore the main issues they focus on are omissions and co-references. Unfortunately, in real-world applications, dialogues are severely noisy and with various kinds of errors that can make their utterance uninformative. Previous benchmarks, as shown in Table 1, either filtered out dialogues with errors (e.g., TASK (Quan et al. 2019), CANARD (Elgohary, Peskov, and Boyd-Graber 2019) and REWRITE (Su et al. 2019)), or ignored this issue and allowed references to still maintain the errors (e.g., Restoration-200K (Pan et al. 2019)). However, the former choice does not match the dialogues from the real world because these errors are almost everywhere in real-world dialogues, while the latter fails to meet the requirements of downstream NLP systems for sentences with high fluency, consistency and coverage.

To this end, this paper presents RealDia, which is constructed from real-world multi-turn dialogues and therefore contains a vast majority of challenges we would face when dealing with real-world dialogues. Our experiments show that the performances of current state-of-the-art dialogue rewriting approaches are dramatically dropped on RealDia, which demonstrates that rewriting real-world dialogues is a challenging task that requires more studies.

Benchmark Construction

Given a dialogue $\boldsymbol{D}$ and an incomplete utterance $\boldsymbol{S}$ in $\boldsymbol{D}$, dialogue rewriting aims to transform S into another sentence $\boldsymbol{S'}$, which can accurately express the speaker's intent of $\boldsymbol{S}$ within the dialogue context. Without loss of generality, we assume that the last utterance in the dialogue is the one to be rewritten, because a dialogue system should be unknown to what the speaker will say in the future. Formally, given $\boldsymbol{D} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_n)$ which is a dialogue containing n utterances, dialogue rewriting aims to generate a dialogue-independent sentence $\boldsymbol{S'} = (x'_1, ..., x'_l)$ such that

$$\text{Intent}(\boldsymbol{S}_n|\boldsymbol{D}) = \text{Intent}(\boldsymbol{S'}). \quad (1)$$

Here Intent() is a function depending on downstream applications, which is difficult to obtain directly. Instead of measuring the intent consistency between $\boldsymbol{S'}$ and $\boldsymbol{S}_n$ in dialogue $\boldsymbol{D}$, we create a golden reference $\boldsymbol{S^*}$ for each case and evaluate the consistency between $\boldsymbol{S'}$ and $\boldsymbol{S}^*$ using automatic and manual evaluation criteria.

We construct Real-world Dialogue Rewriting Corpus (RealDia) from a dialogue corpus provided by a large-scale Chinese Internet company, which contains multi-turn dialogues between users and a widely-used online chatting system in real scenes. Users can interact with the system using both speech or typing inputs, and therefore the (transcribed) dialogues contain various spelling or ASR errors. To ensure the annotation quality, 4 annotators who have degrees in Computational Linguistics are hired to annotate references. Before annotation, we manually collect dialogues whose last utterance cannot clearly express the user intention unless put in the entire dialogue context. Then annotators are asked to create a reference sentence that can fully express the same intention as the last utterance in the dialogue. All references are double-checked to guarantee quality. Finally, RealDia contains 1000 annotated dialogues that need to be rewritten, each of which is manually annotated with a reference. We then randomly sample 700/100/200 dialogues as train/dev/test sets. Besides, we also provide 20,000 dialogues without annotation which can support future research.

To clearly identify the challenges of RealDia, we randomly sampled 200 dialogues and analyze noises appearing in them. We find that in addition to 69.5% of the cases containing NP-ellipsis, there are 40.5% of the cases containing ASR errors or spelling errors. VP-ellipsis, co-references and redundancies are also common. These results show the divergence between RealDia and previous dialogue rewriting benchmarks we discussed, and also demonstrate that RealDia is a far more challenging benchmark.

Skeleton-Guided Dialogue Rewriting

This section describes Skeleton-Guided Rewriter (SGR), which leverages pre-trained text generation models for high fluency and extracts dialogue skeletons for high coverage and consistency to resolve real-world dialogue rewriting. Specifically, as shown in Figure 2, SGR conducts dialogue rewriting with three critical steps: 1) Dialogue Skeleton Extraction, which extracts the critical information across the entire dialogue context to resolve the uninformative utterance problem; 2) Skeleton-Guided Candidate Generation, which generates fluent rewritten candidates under the guidance of extracted skeleton; 3) Skeleton-aware Reranking, which selects the best rewritten sentence by measuring the fluency, coverage and semantic consistency of candidates to the original dialogue. In the following, we will describe these steps and show how each component can be learned with minimum supervision.

Dialogue Skeleton Extraction

Dialogue Skeleton Extractor aims to extract the skeleton which contains the critical information that needs to be covered in the rewritten sentence. For example in Figure 2, given a multi-turn dialogue, Dialogue Skeleton Extractor is expected to extract the critical information "car", "max MPG" and "manufacture" in it. The insight behind Dialogue Skeleton Extractor is to disentangle critical information extraction and sentence generation, so that the critical information in the dialogue history can be better identified to rewrite the uninformative utterance. Furthermore, learning to extract skeletons is much more data-efficient than directly learning to generate rewritten sentences, and therefore much less training data is required for model learning.

Specifically, we formulate skeleton extraction as a token-level sequential labeling problem. Given a dialogue $\boldsymbol{D} = (\boldsymbol{S}_1, ..., \boldsymbol{S}_n)$ and the utterance $\boldsymbol{S}_n$ to be rewritten, we first build the input as:

$$\boldsymbol{D}_E = [\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, [\mathrm{SEP}], \boldsymbol{S}_n], \tag{2}$$

where [SEP] is a special token indicating the start of the last utterance that needs to be rewritten. Then we use RoBERTa (Liu et al. 2019) to encode the input into hidden representations. After that, the representations $\mathbf{h}_j$ for the j^{th} token in $\boldsymbol{D}$ are sent to a binary classifier for classification, where a token with label 1 indicates it is a skeleton token that needs to be covered in the rewritten result, and 0 for otherwise. We combine continuous skeleton tokens into spans, and filtered out stop words to collect dialogue skeleton $\boldsymbol{C} = (c_1, c_2, ..., c_k)$.

Learning. To train the Dialogue Skeleton Extractor, we label tokens in a dialogue by comparing the dialogue $\boldsymbol{D}$ with its golden reference $\boldsymbol{S}^*$, i.e., the ith token will be labeled as $y_i = 1$ if it appears in $\boldsymbol{S}^*$ and 0 otherwise. Based on the token labels, we train the Dialogue Skeleton Extractor by optimizing the cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{n} [y_i \cdot \log P_i + (1 - y_i) \cdot \log (1 - P_i)], \tag{3}$$

where P_i is the predicted probability of the ith token as a skeleton token.

Skeleton-Guided Candidate Generation

Given a dialogue $\boldsymbol{D} = (\boldsymbol{S}_1, ..., \boldsymbol{S}_n)$ and its skeleton spans $\boldsymbol{C} = (c_1, c_2, ..., c_k)$, Skeleton-Guided Generator will generate rewritten candidates $\boldsymbol{S}' = (x_1', \cdots, x_l')$ under the guidance of skeleton $\boldsymbol{C}$. Because current large-scale pre-trained generation models like T5 (Raffel et al. 2020) can effectively generate fluent natural language sentences, we build Skeleton-Guided Generator by directly leveraging the T5-based encoder-decoder architecture and further incorporating skeleton to improve the coverage and consistency of model outputs. Inspired by recent advances in prompt mechanism for text generation (Brown et al. 2020; Zou et al. 2021) and skeleton-based generation models (Xu et al. 2018; Cai et al. 2019; Su et al. 2021), we transform skeleton into prompt-style guidance, which is expected to guide the generator to pay more attention to critical information in the dialogue to resolve uninformative utterance challenge.

Formally, given a dialogue $\boldsymbol{D}$ and its skeleton $\boldsymbol{C} = (c_1, c_2, ..., c_k)$, the input of Skeleton-Guided Generator is:

$$\boldsymbol{D}_G = [c_1, [\mathrm{SEP}], ..., c_k, [\mathrm{CLS}], \boldsymbol{D}] \tag{4}$$

where [SEP] is used for segmentation and [CLS] indicates the beginning of a dialogue. Then $\boldsymbol{D}_G$ is fed into a T5-based encoder-decoder architecture to generate rewritten sentence $\boldsymbol{S}'$ as:

$$P(\boldsymbol{S}' \mid \boldsymbol{D}_G) = \prod_t P\left(y_t' \mid \boldsymbol{D}_G, y_{<t}'\right), \tag{5}$$

where y_t' is the token generated at time step t, and $y_{<t}'$ is the generated result before y_t'. To improve the diversity of candidate sentences and reduce the error propagation impact of wrong skeleton spans, we form different skeletons by randomly sampling N skeleton spans from $\boldsymbol{C}$ several times. Then we preserve Top-3 candidates for each skeleton combination until we obtain a pre-defined number of candidates.

Skeleton-aware Pre-training. To ensure that the generator focuses on the skeleton, we further pre-train Skeleton-Guided Generator from T5 by automatically constructing a pseudo dataset. We first use natural language inference model to collect data from unlabeled dialogues. For each collected dialogue $\boldsymbol{D} = \{\boldsymbol{S}_1, ..., \boldsymbol{S}_n\}$, its last utterance $\boldsymbol{S}_n$ entails the critical information of its previous dialogue history $\{\boldsymbol{S}_1, ..., \boldsymbol{S}_{n-1}\}$. Then for each collected dialogue $\boldsymbol{D}$, we construct a pseudo instance $(\boldsymbol{D} - \boldsymbol{S}_n, \boldsymbol{S}_{n-1}, \boldsymbol{S}_n)$, where $\boldsymbol{D} - \boldsymbol{S}_n$ is regarded as the dialogue, $\boldsymbol{S}_{n-1}$ is regarded as the utterance to rewrite and $\boldsymbol{S}_n$ is regarded as the golden reference. Finally, we use the verbs and nouns both in $\boldsymbol{S}_n$ and $\boldsymbol{D}$ as skeleton spans $\boldsymbol{C}$ to pre-train Skeleton-Guided Generator using these pseudo instances.

After pre-training, Skeleton-Guided Generator is further fine-tuned using labeled data by optimizing the following likelihood-based training objective:

$$\mathcal{L} = -\sum_{t=1}^{l} \log P\left(x_t \mid x_{<t}; \boldsymbol{D} - \boldsymbol{S}_n, \boldsymbol{C}, \boldsymbol{\theta}\right) \tag{6}$$

Skeleton-aware Reranking

Pre-trained text generation models can effectively generate highly fluent texts. Unfortunately, they will occasionally generate text that is nonsensical, or unfaithful to the provided source input, which is referred as hallucination (Raunak, Menezes, and Junczys-Dowmunt 2021; Dziri et al. 2021; Ji et al. 2022). However, dialogue rewriting requires the rewritten sentences to be not only fluent, but also with high coverage and high consistency to the original dialogue. In order to ensure the quality of generated sentences, this paper further proposes a skeleton-aware reranker, which reranks candidates based on three measurements which can effectively reflect the fluency, consistency and coverage of rewritten sentences.

Fluency Measurement. Because large-scale pre-trained language models like T5 (Raffel et al. 2020) can generate fluent natural language sentences, this paper directly uses the generation probability of a candidate to measure its fluency $f_{Flu}(\boldsymbol{S'}|\boldsymbol{D}, \boldsymbol{C})$.

Consistency Measurement. Because PLMs can frequently generate context-inconsistent sentences, we design a consistency measurement to penalize outputs containing context-irrelevant information. Specifically, we observe that the extracted skeleton spans can effectively cover the critical information that needs to appear in the rewritten sentence. Therefore, to measure the consistency between candidates and the original dialogue context, we count the number of non-stop words in each candidate that are not in the skeleton span set. The negative value of the number of irrelevant words in each candidate is used as the consistency measurement $f_{Con}(\boldsymbol{S'}, \boldsymbol{C})$.

Coverage Measurement. To estimate how well the rewritten sentence covers the critical information of the dialogue, we introduce a coverage measurement based on natural language inference (NLI), i.e., estimate whether the rewritten sentence can entail the original dialogue. Specifically, we regard the candidate as the premise and the original dialogue as the hypothesis of the NLI task. We concatenate the candidate and dialogue history with [SEP] and feed it into a RoBERTa-based NLI model, then the NLI probability of entailment relation is used as the coverage measurement $f_{Cov}(\boldsymbol{D}, \boldsymbol{S'}|\boldsymbol{C})$.

The Reranker. The final skeleton-aware reranker is a linear combination of the above three measurements:

$$\mathrm{R} = \lambda_1 f_{Flu}(\boldsymbol{S'}|\boldsymbol{D}, \boldsymbol{C}) + \lambda_2 f_{Con}(\boldsymbol{S'}, \boldsymbol{C}) + \lambda_3 f_{Cov}(\boldsymbol{D}, \boldsymbol{S'}|\boldsymbol{C}),$$

where $\lambda_1 + \lambda_2 + \lambda_3 = 1$, which represent the weights of the three measurements respectively. In this paper, we find the best weights using grid search on the development set with the step size set to 0.05.

Experiments

Experiment Setup

Baselines. We compare the following baselines with SGR.

- **T-Ptr-λ** (Su et al. 2019) is a pure pointer-based rewriter based on transformers, which can only copy the word from the input.
- **PAC** (Pan et al. 2019) is a pick-and-combine model to first pick omitted words from context and then combine them with the incomplete utterance.
- **RUN** (Liu et al. 2020) formulates incomplete utterance rewriting as a semantic segmentation task. Here we use the BERT+RUN model.
- **RAST** (Hao et al. 2021) treats utterance rewriting as multi-task sequence tagging, and injects the loss signal from BLEU or GPT-2 under a reinforce framework to improve the fluency. Here we use the BERT-Large+RAST.
- **JET** (Inoue et al. 2022) jointly optimizes picking tokens and generating rewritten utterances based on T5.
- **T5** (Raffel et al. 2020) introduces a unified framework that converts NLP tasks into a text-to-text format.

Implementation. Our implementation is based on PyTorch (Paszke et al. 2019) and the Transformers library of HuggingFace (Wolf et al. 2020). We build Skeleton-Guided Generator based on pre-trained Chinese T5-base[1], and construct Dialogue Skeleton Extractor and NLI model based on pre-trained Chinese RoBERTa (Cui et al. 2021). For the weights of three features in Reranker, we take λ_1=0.3, λ_2=0.2, and λ_3=0.5, which are the best weights on the development set. We optimized our model using label smoothing (Szegedy et al. 2016; Müller, Kornblith, and Hinton 2019) and AdamW (Loshchilov and Hutter 2019) with the learning rate=1e-4. In the experiments, each compared model is trained for 30 epochs with a batch size of 16.

Evaluation Metrics

We use the widely-used metrics including BLEU (Papineni et al. 2002), ROUGE (Lin and Hovy 2002), EM (exact match), and BERTScore (Zhang et al. 2020) as the automatic evaluation metrics. In addition, we also hire 4 professional data annotators to evaluate the quality of model outputs. Specifically, a rewritten sentence is recorded as 1 if it is fluent and semantically equivalent to the target utterance in the original dialogue, otherwise is 0. For ablation studies, we also perform a more fine-grained human evaluation from three aspects: Fluency, Coverage, and Consistency. Fluency evaluates the quality of each generated sentence, Consistency evaluates the factual alignment between the original dialogue and rewritten sentence, and Coverage measures how well the rewritten sentence covers the key information of the dialogue that needs to appear in the target output. We ask human annotators to rate each output on the scale of [1, 5] (higher is better). We average the scores of different annotators as the final performance.

Overall Results on RealDia

Table 2 shows the results of our model and baselines on RealDia. From this table, we can see that:

1. Real-world dialogue rewriting is very challenging due to the existence of various kinds of errors and noises. As a result, previous state-of-the-art approaches can not achieve reasonable performance on RealDia. Although

[1]https://github.com/ZhuiyiTechnology/t5-pegasus

	BLEU-2	BLEU-4	ROUGE-2	ROUGE-L	BERTScore-F	EM	Human Evaluation
T-Ptr-λ	38.7	18.4	30.3	60.4	73.6	0	0
PAC	62.2	48.7	63.7	81.7	86.5	15.0	20.0
BERT+RUN†	70.7	59.0	71.8	82.1	89.0	20.0	36.0
BERT-L+RAST†	79.9	76.3	75.5	82.8	90.1	31.5	47.0
JET	81.9	73.7	81.4	88.4	93.3	38.0	50.5
T5-base	78.6	70.6	78.6	86.8	93.0	40.0	62.5
SGR(Ours)	**83.9**	**76.4**	**83.4**	**89.2**	**93.8**	**49.5**	**72.5**

Table 2: Automatic and Human Evaluation results of different models on the RealDia test set. Results marked with † are from our runs with their released code. Our model significantly outperforms other models on all metrics.

Model	EM	B_2	B_4	R_2	R_L
T-Ptr-Net	53.0	83.9	77.1	85.1	88.7
T-Ptr-Gen	53.1	84.4	77.6	85.0	89.1
T-Ptr-λ	52.6	85.6	78.1	85.0	89.0
RUN	53.8	86.1	79.4	85.1	89.5
T5-base	67.3	89.5	84.1	90.1	93.4
BERT + T-Ptr-λ	57.5	86.5	79.9	86.9	90.5
BERT + RUN	66.4	**91.4**	86.2	90.4	93.5
BERT + RAST	64.3	90.2	**88.2**	88.9	91.5
JET	69.1	91.2	86.6	90.6	-
SGR (Ours)	**69.4**	90.6	85.4	**90.9**	**93.8**

Table 3: Experiment results on REWRITE. B_i represents the BLEU-i score and R_i is the ROUGE-i score. EM is the exact match metric. The results of T-Ptr and RUN baselines are adopted from Liu et al. (2020). The results of BERT + RAST and JET are adopted from Hao et al. (2021) and Inoue et al. (2022) respectively.

baseline approaches have been proven effective on previous dialogue rewriting benchmarks shown in Table 3, their performances are all significantly degraded on RealDia. We believe this is because previous methods are built upon the informative utterance assumption, and therefore they focus on how to directly complete original utterances. Unfortunately, uninformative utterances to rewrite are widely spread in real-world dialogues due to the existence of noises and errors. As a result, previous approaches are unable to handle these issues. This result demonstrates the necessity of RealDia and SGR.

2. SGR can effectively resolve the uninformative utterance challenge via skeleton-guided generation and therefore achieve remarkable improvement on RealDia. Compared with previous baselines, SGR achieves consistent and significant improvements on both Automatic Evaluation and Human Evaluation, which outperforms the previous best dialogue rewriting model JET by 2.7 points on BLEU-4 and 2.0 points on ROUGE-2. Furthermore, we can see that its improvements in Human Evaluation and EM are far more significant. This demonstrates SGR can not only generate better rewritten sentences in surface form, but also can better meet the requirements of downstream NLP applications.

3. Skeleton is critical for consistent and high coverage rewritten sentence generation. Compared with original T5 model without skeleton guidance, SGR can improve 5.8 points on BLEU-4 and 9.5 points on EM, as well as 10 points on Human Evaluation. We believe this is because SGR can exploit training data more efficiently by focusing on the consistency and coverage of rewritten sentences but still taking advantage of the high fluency of large-scale pretrained language model outputs.

	Automatic Evaluation			Human Evaluation		
	B_4	R_L	EM	Flu.	Cov.	Con.
SGR	**76.4**	**89.2**	**49.5**	**4.87**	**4.57**	**4.39**
w/o Skeleton	74.8	88.3	45.0	4.74	4.37	4.15
w/o Rerank	74.1	88.1	45.0	4.85	4.46	4.27
w/o Pretrain	73.7	88.2	43.5	4.85	4.44	4.22
T5	70.6	86.8	40.0	4.78	4.33	4.08

Table 4: Ablation results in Automatic and Human Evaluation on RealDia. Flu. represents Fluency, Cov. represents coverage and Con. represents consistency.

Experiment Results on REWRITE

To demonstrate the effectiveness of SGR, we also conduct experiments on REWRITE, a representative dialogue rewriting dataset. Table 3 shows the results of SGR and previous state-of-the-art models on REWRITE. We can see that SGR can effectively resolve omissions and co-references in REWRITE benchmark, and achieve state-of-the-art performance. Besides, we can see that the advantage of SGR is more obvious in exact match metric compared to other automatic evaluation metrics. We believe this is because SGR is guided to generate rewritten sentences with high fluency, consistency and coverage, rather than directly editing the original utterance. As a result, the automatic evaluation metrics may under-estimate the improvements of SGR, which is also consistent with the results in Table 2.

Ablation Studies

To analyze the effects of skeleton-aware pre-training, skeleton guidance, and skeleton-aware reranker, we conduct ablation studies on RealDia, and the results are shown in Table 4. We can see that: 1) Skeleton is critical for rewriting multi-turn dialogues. After removing the guidance of skeleton (denoted as -Skeleton), the Fluency, Consistency, and Coverage

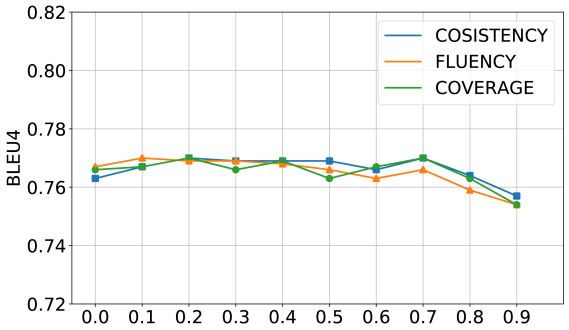

Figure 3: Best BLEU-4 SGR achieved when Fluency, Coverage, and Consistency feature weights are fixed in turn.

scores all drop on both Automatic Evaluation and Human Evaluation. 2) Skeleton-aware Reranker can effectively select candidates with high fluency, coverage, and semantic consistency. After removing the Skeleton-aware Reranker of SGR (denoted as -Rerank), the scores of Consistency and Coverage all decrease by more than 2%, but it does not have much impact on Fluency. T5 can generate high fluency sentences but its output with the highest generation probability cannot ensure the coverage and consistency of their outputs. 3) Skeleton-aware pre-training can effectively improve the consistency and coverage of rewritten sentences. We can see that Skeleton-aware pre-training improves the scores of Coverage and Consistency by 2.9% and 4.0%, respectively.

Effects of Different Reranking Measurements

To explore the effects of Fluency, Coverage, and Consistency features in Reranker, we fix each feature weight to take values from 0.1 to 0.9 in turn and record the best performance that the model can achieve when that feature weight takes the corresponding value. The step is 0.1, and we ensure that the sum of the three feature weights is always 1. The BLEU-4 results are the average of three replicate runs, as shown in Figure 3. The best results of fixing different feature weights have approximately the same trend on BLEU-4. First, the effect of fixing one feature weight is relatively small, with the range being less than 0.02 on BLEU-4, demonstrating our architecture's robustness to some extent. Second, the performance of the model is relatively poor when the value of a feature weight is too small ($<$0.2) or too large ($>$0.7), which indicates that all three features are crucial and need to be considered together.

Case Analysis

To demonstrate and compare the effect of different methods, Figure 4 shows two cases that contain the outputs of different systems. We can see that:

1. RealDia is challenging for previous dialogue rewriting approaches. First, baselines frequently generate illegal sentences with low fluency and grammar errors, such as omissions, repetitions, and word order errors. In these cases, the outputs of T-Ptr-λ contain many word order errors, and PAC gets stuck in repetitive loops. In addition, in case 1, T-Ptr-λ, PAC, and RAST generate illegal sentences, and the

	Case1	Case2
Dialogue (Translation)	S_1 :怎么记录(jì lù)一只小猫咪 (How to record a kitten) S_2: 怎么激怒(jī nù) (How to irritate)	S_1: 中国象棋的队 (Chinese chess team) S_2: 双方各有几种棋子 (How many kinds of pieces are there on both sides)
Reference (Translation)	S^*: 怎么激怒(jī nù)一只小猫咪 (How to irritate a kitten)	S^*: 中国象棋双方各有几种棋子 (How many kinds of pieces are there on both sides of Chinese chess)
T-Ptr-λ (Translation)	S': 怎么猫怎么怒激激(nù jī jī) (How the kitten how #error)	S': 中国国的中各几种棋子 (How many kinds of chess pieces are there in China #error)
PAC (Translation)	S': 怎么小猫咪怎么激怒激怒(jī nù jī nù) (How the kitten how to irritated #repetition)	S': 中国象棋双方各几种棋子子子子子 (How many kinds of pieces #repetition are there on both sides of Chinese chess)
BERT+RUN (Translation)	S': 一只怎么激怒(jī nù) (One #ellipsis how to irritate)	S': 中国象棋的双方各有几种棋子 (How many kinds of pieces are there on both sides of Chinese chess)
BERT-L+RAST (Translation)	S': 怎么记录(jì lù)小猫咪怎么激怒(jī nù) (How to record a kitten how to irritate)	S': 中国象棋双方各有几种棋子 (How many kinds of pieces are there on both sides of Chinese chess)
JET (Translation)	S': 怎么记录(jì lù)一只猫咪 (How to record a kitten)	S': 中国象棋的各有几种棋子 (How many kinds of Chinese chess pieces are there on each side of #ellipsis)
T5 (Translation)	S': 怎么记录(jì lù)一只猫咪的表情 (How to record the expression of a kitten)	S': 中国象棋的三大队各有几种棋子 (How many kinds of Chinese chess pieces are there in each of the three teams)
SGR(Ours) (Translation)	S': 怎么激怒(jī nù)一只小猫咪 (How to irritate a kitten)	S': 中国象棋的两队各有几种棋子 (How many kinds of pieces are there on both sides of Chinese chess)

Figure 4: Two example dialogues and their references from the RealDia, as well as the rewritten sentences generated by different models. We mark errors in red.

subject has been omitted from the output of RUN. Second, it is hard for baselines to distinguish between noises and critical information in dialogues due to their inherent informative utterance assumption. RAST, T5 and JET fail to correct the ASR error in original dialogue in case 1, while SGR can correct the transcription error and generate fluent utterances.

2. Compared with previous state-of-the-art approaches, T5 can guarantee the fluency of its outputs but may have consistency and coverage issues. Under the guidance of skeleton, SGR can improve the consistency and coverage of rewritten sentences. In these cases, the outputs of T5 contain hallucination contents "expression" and "three teams" that do not appear in the dialogue history, while SGR can correctly generate fluent utterances with high consistency and high coverage based on the dialogue context.

Conclusion

Previous dialogue rewriting benchmarks and systems assume a fluent and informative utterance to rewrite, which is inconsistent with real-world dialogue rewriting application scenarios. In this paper, we first present Real-world Dialogue Rewriting Corpus (RealDia), a new benchmark to evaluate how well current dialogue rewriting systems can deal with real-world noisy and uninformative dialogue utterances. Then we present Skeleton-Guided Rewriter (SGR), which can resolve the task via a skeleton-guided generation paradigm. Experiments on RealDia and previous benchmarks have shown that rewriting real-world noisy dialogues is challenging, and SGR achieves state-of-the-art performance on both RealDia and previous benchmarks.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work is supported by the Natural Science Foundation of China (No.U1936207, 62122077, 62106251 and 61906182).

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*.

Cai, D.; Wang, Y.; Bi, W.; Tu, Z.; Liu, X.; Lam, W.; and Shi, S. 2019. Skeleton-to-Response: Dialogue Generation Guided by Retrieval Memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 1219–1228. Minneapolis, Minnesota: Association for Computational Linguistics.

Carbonell, J. G. 1983. Discourse Pragmatics and Ellipsis Resolution in Task-Oriented Natural Language Interfaces. In *21st Annual Meeting of the Association for Computational Linguistics*, 164–168. Cambridge, Massachusetts, USA: Association for Computational Linguistics.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z.; Wang, S.; and Hu, G. 2021. Pre-Training With Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3504–3514.

Dziri, N.; Madotto, A.; Zaïane, O.; and Bose, A. J. 2021. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2197–2214. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Elgohary, A.; Peskov, D.; and Boyd-Graber, J. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5918–5924. Hong Kong, China: Association for Computational Linguistics.

Hao, J.; Song, L.; Wang, L.; Xu, K.; Tu, Z.; and Yu, D. 2021. RAST: Domain-Robust Dialogue Rewriting as Sequence Tagging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4913–4924. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Inoue, S.; Liu, T.; Nguyen, S.; and Nguyen, M.-T. 2022. Enhance Incomplete Utterance Restoration by Joint Learning Token Extraction and Text Generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3149–3158. Seattle, United States: Association for Computational Linguistics.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; and Fung, P. 2022. Survey of Hallucination in Natural Language Generation. *CoRR*, abs/2202.03629.

Jin, L.; Song, L.; Jin, L.; Yu, D.; and Gildea, D. 2022. Hierarchical Context Tagging for Utterance Rewriting. In *AAAI*.

Lin, C.-Y.; and Hovy, E. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, 45–51. Phildadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Liu, Q.; Chen, B.; Lou, J.-G.; Zhou, B.; and Zhang, D. 2020. Incomplete Utterance Rewriting as Semantic Segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2846–2857. Online: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations*. OpenReview.net.

Mele, I.; Muntean, C. I.; Nardini, F. M.; Perego, R.; Tonellotto, N.; and Frieder, O. 2021. Adaptive utterance rewriting for conversational search. *Information Processing & Management*, 58(6): 102682.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, 4696–4705.

Pan, Z.; Bai, K.; Wang, Y.; Zhou, L.; and Liu, X. 2019. Improving Open-Domain Dialogue Systems via Multi-Turn Incomplete Utterance Restoration. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 1824–1833. Hong Kong, China: Association for Computational Linguistics.

Pangaro, P.; and Dubberly, H. 2014. What is Conversation? How Can We Design for Effective Conversation? In *Driving Desired Futures*, 144–159. Birkhäuser.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, 8024–8035.

Quan, J.; Xiong, D.; Webber, B.; and Hu, C. 2019. GECOR: An End-to-End Generative Ellipsis and Co-reference Resolution Model for Task-Oriented Dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 4547–4557. Hong Kong, China: Association for Computational Linguistics.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Raunak, V.; Menezes, A.; and Junczys-Dowmunt, M. 2021. The Curious Case of Hallucinations in Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1172–1183. Online: Association for Computational Linguistics.

Si, S.; Zeng, S.; and Chang, B. 2022. Mining Clues from Incomplete Utterance: A Query-enhanced Network for Incomplete Utterance Rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4839–4847. Seattle, United States: Association for Computational Linguistics.

Su, H.; Shen, X.; Zhang, R.; Sun, F.; Hu, P.; Niu, C.; and Zhou, J. 2019. Improving Multi-turn Dialogue Modelling with Utterance ReWriter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 22–31. Florence, Italy: Association for Computational Linguistics.

Su, Y.; Wang, Y.; Cai, D.; Baker, S.; Korhonen, A.; and Collier, N. 2021. PROTOTYPE-TO-STYLE: Dialogue Generation With Style-Aware Editing on Retrieval Memory. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 2152–2161.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826. IEEE Computer Society.

Wang, X.; Zhang, H.; Zhao, S.; Zou, Y.; Chen, H.; Ding, Z.; Cheng, B.; and Lan, Y. 2021. FCM: A Fine-grained Comparison Model for Multi-turn Dialogue Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4284–4293. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Wu, Y.; Wu, W.; Xing, C.; Zhou, M.; and Li, Z. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 496–505. Vancouver, Canada: Association for Computational Linguistics.

Xu, J.; Ren, X.; Zhang, Y.; Zeng, Q.; Cai, X.; and Sun, X. 2018. A Skeleton-Based Model for Promoting Coherence Among Sentences in Narrative Story Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4306–4315. Brussels, Belgium: Association for Computational Linguistics.

Zhang, R.; Yu, T.; Er, H.; Shim, S.; Xue, E.; Lin, X. V.; Shi, T.; Xiong, C.; Socher, R.; and Radev, D. 2019. Editing-Based SQL Query Generation for Cross-Domain Context-Dependent Questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5338–5349. Hong Kong, China: Association for Computational Linguistics.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations*. OpenReview.net.

Zou, X.; Yin, D.; Zhong, Q.; Yang, H.; Yang, Z.; and Tang, J. 2021. Controllable Generation from Pre-Trained Language Models via Inverse Prompting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, 2450–2460. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383325.