Controllable Conversation Generation with Conversation Structures via Diffusion Models

Jiaao Chen
Georgia Institute of Technology
jchen896@gatech.edu

Diyi Yang
Stanford University
diyiy@cs.stanford.edu

Abstract

Generating coherent conversation is an important and challenging long text generation task, as it has various applications such as daily entertainment, education, or building conversational AI to facilitate human-computer interaction. However, current generation models often fail to effectively utilize rich linguistic and world knowledge to generate conversations just like humans. In this work, we introduce a novel conversation generation framework to effectively incorporate human knowledge and conversation structures with both controllability and interpretability for better conversation generation. Specifically, we first generate the prototype conversations from short descriptions. We then gradually and strategically incorporate different levels of conversation structures including the *action triples*, *dialogue acts*, and *discourse relations* via diffusion models to directly edit the prototype conversations. We demonstrate the effectiveness of our framework through experiments on two datasets by comparing our method with the state-of-the-art baseline models[1].

1 Introduction

Generating long-form and coherent text is an important step in many natural language generation (NLG) applications (Guan et al., 2022). While recent research has shown impressive progress in generating short texts, it is still challenging for generation models to write coherent long text which requires comprehensively incorporating linguistic and world knowledge (Charniak, 1972). Our work takes a closer look at long conversation generation (Gunasekara et al., 2021), one of the most challenging long text generation tasks. The task is to generate an entire coherent conversation from a given short description, i.e., a summary, of it. Conversation generation has various applications from daily entertainment, and story generation, to customer services. However, real human/human conversation logs are scarce; crowdsourcing conversational data is time-consuming, costly, and hard to ensure data quality (Gunasekara et al., 2021). Thus, better conversation generation models would allow us to generate massive natural conversational data more automatically and efficiently, which further helps build better conversational AI systems.

Even though there are a growing number of studies that focused on long text generation such as story generation (Guan et al., 2022; Yang et al., 2022; Fan et al., 2018; Li et al., 2022a) using large pre-trained models (Fan et al., 2018; Yang et al., 2022), event planning (Guan et al., 2022; Fan et al., 2018; Li et al., 2022a) and recursive revision (Yang et al., 2022), directly applying them to generate long conversation may not work well due to the inherent different structures between stories and conversations. For instance, previous long text generation usually focused on generating stories that talk about one single topic with five sentences to one paragraph. They are shorter compared to conversations, which usually cover multiple topics between different speakers (over ten turns) (Feng et al., 2020). Furthermore, there are diverse discourse relations between different speakers (Chen and Yang, 2021b), making it even more challenging to generate long and coherent conversations.

While there is a line of work about dialogue generation, they are mainly concentrated on generating the next utterance autoregressively based on the given context (Ji et al., 2021; Liu et al., 2020; Saha et al., 2022; Zhang et al., 2020; Ramakrishnan et al., 2022) with sequence-to-sequence models. Such methods usually neglect the conversation structures(Adewumi et al., 2022), and thus might easily lose focus to produce long and coherent conversations after several rounds of generation (Gunasekara et al., 2021). Moreover, the formerly generated utterances could not be further edited to

[1] The code is available at https://github.com/SALT-NLP/Conversation_Generation_Diffusion

Findings of the Association for Computational Linguistics: ACL 2023, pages 7238–7251
July 9-14, 2023 ©2023 Association for Computational Linguistics

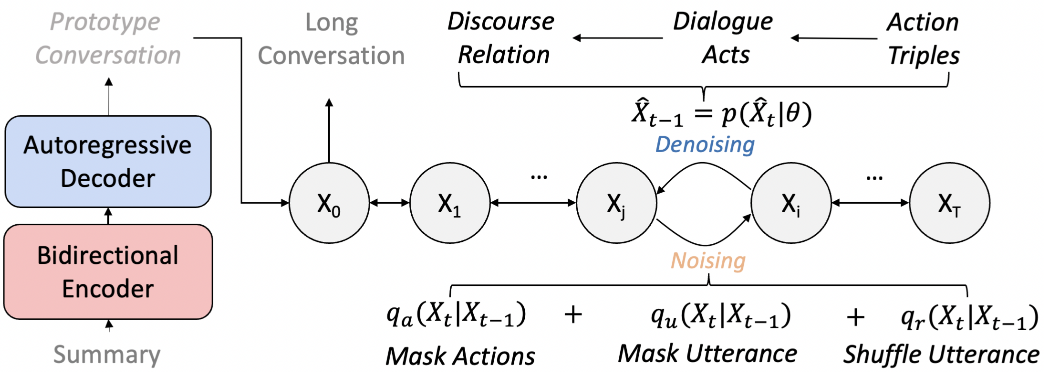

Figure 1: Overall process of our framework. The sequence-to-sequence models would first generate a prototype conversation. Then we first corrupt the conversation through masking actions, masking utterances, shuffling utterances in the forward process, and utilize the diffusion process to gradually enrich the prototype conversation with different levels of structured conversation information.

adapt the later generated utterances. It is also unclear whether and how these sequence-to-sequence models are "gradually planning" to produce the long conversations. Therefore, how to design controllable methods tailored to the structures in conversations for generating long and coherent conversations becomes especially important.

To this end, our work introduces a Controllable Conversation Generation Framework with Diffusion Models (**Diffuse-CG**, shown in Figure 1) to incorporate different conversational structures in a *non-autoregressive* manner, inspired by recent advances in deep generative models (Li et al., 2022b; Gong et al., 2022; He et al., 2022). Specifically, we first generate a prototype conversation using pre-trained sequence-to-sequence model based on the input description. Then we leverage the diffusion models to gradually enrich the prototype conversation with conversation structures. The diffusion process allows a more flexible conversation generation by not limiting a fixed left-to-right generation order; it also allows the model to gradually incorporate different levels of conversation structures to control the granularities, including the use of *action triples* to add more specific topics and events (Gee, 2014; Chen and Yang, 2021b), *dialogue acts* to make the utterances more like human (Allen and Core, 1997; Sacks et al., 1978; Chen and Yang, 2021a), and *discourse relations* to generate longer conversations with better coherency (Kirschner et al., 2012; Stone et al., 2013; Asher et al., 2016a). To make the diffusion process more adapted to conversation generation and more stable, we further improve the general diffusion model (Li et al., 2022b) with *linguistic-informed noise* where we perturb the prototype conversation in the forward process with noise including soft-masking action words, soft-masking utterances, and shuffling discourse relations, rather than pure Gaussian noise (Li et al., 2022b). Experiments on two conversation datasets, SAMSum (Gliwa et al., 2019) and DialogSum (Chen et al., 2021)by visualizing the intermediate-generated conversations, we show that Diffuse-CG achieves better interpretability for understanding how the model is structuring and generating long-form conversations.

2 Related Work

Long Text Generation Long-form text generation has been a longstanding challenge in natural language generation where models need to generate long, coherent and open-ended narratives (Guan et al., 2022; Yang et al., 2022; Fan et al., 2018; Li et al., 2022a; Guan et al., 2021). Recent studies have shown impressive success in generating more coherent stories through adopting hierarchical model structures (Li et al., 2015), leveraging large pre-trained models (Fan et al., 2018; Yang et al., 2022), planing first and then generating framework (Shao et al., 2019; Tan et al., 2020; Goldfarb-Tarrant et al., 2020; Li et al., 2022a) and incorporating external knowledge (Guan et al., 2022; Fan et al., 2018; Xu et al., 2020). However, previous studies mainly focus on generating single-speaker stories and neglect one important form of long text—conversations. Such methods cannot be directly applied to generate multi-speaker conversations because of the complex linguistic structures in conversations such as back-and-forth interactions (Feng et al., 2020; Chen and Yang, 2021b).

Our work fills this gap by utilizing conversation structures to generate coherent conversations.

Dialogue Response Generation Numerous studies have been conducted on generating short responses conditioned on previous context (Ji et al., 2021; Liu et al., 2020; Saha et al., 2022; Zhang et al., 2020; Ramakrishnan et al., 2022) such as adding user's persona (Wolf et al., 2019), paraphrasing template responses (Lippe et al., 2020) and using example guidance (Gupta et al., 2021; Cai et al., 2020). While achieving state-of-the-art performances, they suffer from generating the entire conversation because they can only generate one utterance at a time and easily lose focus when generating multiple rounds of utterances or the entire conversation (Gunasekara et al., 2021). This is largely due to the fact that former errors cannot be corrected when generating utterance by utterance autoregressively, and the lack of awareness towards rich conversation structures like long-distance relations in conversations (Stone et al., 2013; Asher et al., 2016a). To this end, we design a controllable and interpretable conversation generation framework that makes use of rich structures to generate the entire conversation in a non-autoregressive way.

Diffusion Model Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) are recently-introduced state-of-the-art non-autoregressive generative models and have shown substantial success for visual modalities (Ramesh et al., 2022; Rombach et al., 2022). They are generally more interpretable and controllable as they gradually denoise random vectors to desired output via multiple intermediate steps (He et al., 2022; Austin et al., 2021). However, it is still difficult to apply diffusion models to textual data, because the input space in text is discrete and text is generally more complex in structures. Although there are a few exceptions to model language generation with diffusion process (Li et al., 2022b; Gong et al., 2022; He et al., 2022; Austin et al., 2021; Hoogeboom et al., 2021) where continuous and discrete space is bridged through embedding and rounding (Li et al., 2022b; Gong et al., 2022; Dieleman et al., 2022), such approaches often utilize Gaussian noise in the forward process, which usually fails to leverage the linguistic structure in text to noise the input textual data and makes the diffusion models unstable and costly (He et al., 2022). Building upon these prior works, we utilize diffusion models for interpretable and controllable conversation generation and design a novel *linguistic-informed noise* for adapting diffusion models to generate textual conversations.

3 Background: Diffusion Models

Diffusion models are the recent state-of-the-art deep generative models via iterative denoising the latent variables (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). Basically, corruption (usually Gaussian noise) is added to the input data distribution gradually during a forward process. Then a diffusion model is trained through learning to recover the corrupted distribution to the original input data distribution step by step. A small amount of information that is perturbed during the corresponding forward process is reconstructed in every diffusion step.

There is usually a forward noising process and a diffusion denoising process in a diffusion model. For a given sampled input data, $x_0 \sim q(x_0)$, a Markov chain of latent variables $\{x_1, \cdot\cdot\cdot, x_T\}$ are generated in the forward noising process $(q\left(x_t \mid x_{t-1}\right))$ by progressively adding a small amount of Gaussian noise to perturb the input data:

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right)$$

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ is a noise schedule controlling the amount of added noise in every step. Through the forward process, x_T becomes an isotropic Gaussian distribution. Note that there are no trainable parameters in the forward process.

Then a reversed diffusion process, which is learned by a parameterized model $(p(x_{t-1}|x_t))$, is learned to denoise x_T to the original data x_0:

$$p_\theta\left(x_{t-1} \mid x_t, t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right),$$

where $\mu_\theta(.)$ and $\Sigma_\theta(.)$ are the learned models.

The diffusion model is trained to maximize the marginal likelihood of $\log p_\theta(x_0)$. And Ho et al. expand and reweight the objectives to obtain a mean-squared error (L_2) loss:

$$\mathcal{L}_{\mathrm{d}}\left(x_0\right) = \sum_{t=1}^{T} \underset{q(x_t|x_0)}{\mathbb{E}} \left\|\mu_\theta\left(x_t, t\right) - \hat{\mu}\left(x_t, x_0\right)\right\|^2$$

where $\hat{\mu}$ is the mean of the posterior $q(x_{t-1}|x_0, x_t)$, and μ_θ is the predicted mean of $p_\theta(x_{t-1}|x_t)$, which is predicted by the parameterized neural models.

4 Our Approach

This section introduces our controllable conversation generation model to generate natural and coherent conversations, as shown in Figure 1. Basically, we first utilize a sequence-to-sequence model to generate a prototype version of the conversation based on the given short description (Section 4.1). We then gradually incorporate the conversation structure guidance to edit the prototype conversation in order from lower levels to higher levels (action triples, dialogue acts, and discourse relations) through diffusion models (Section 4.2).

4.1 Prototype Conversation Generation

We first train a sequence-to-sequence model $f(F(.))$ to generate the prototype conversation C_p based on the given conversation summary s, $C_p = f(F(s))$, where $F(.)$ is an encoder-decoder network and $f(.)$ is a feed-forward network to map the hidden representations to actual words. We initialize $f(F(.))$ with a pre-trained encoder-decoder model, i.e., BART-base (Lewis et al., 2020). $f(F(.))$ is learned using the ground truth summary-conversation pairs, (s, C_g) through minimizing the cross entropy $\mathcal{L} = -\sum \log P(C_g|s)$.

Once the prototype conversation generation model is learned, we utilize $F(.)$ to generate the hidden representations $X_0 = \{w_0, ..., w_l\}$ of the prototype conversation C with l words: $X_0 = \{w_0, ..., w_l\} = F(s)$. Note that $X_0 \in R^{l \times d}$ is a matrix used as the initiate latent variable in Section 4.2, where l is the number of words in the conversation and d is the dimension of the hidden representation.

4.2 Editing with Diffusion Models

With the hidden representation, X_0, of the prototype conversation, we then introduce our diffusion model that gradually edits the prototype conversation to form the desired long conversation. Specifically, we first add linguistic noise to X_0 to get the noisy intermediate latent variables $X_{1:T}$ in the forward process (Section 4.2.2), and then gradually denoise X_T to $\hat{X}_0$ with different levels of conversation structure information in the diffusion process (Section 4.2.3). Last, we generate the long conversation C_l with the denoised $\hat{X}_0$: $C_l = f(\hat{X}_0)$.

4.2.1 Structures in Conversations

This part introduces the three types of widely-used structures with different granularity in conversations utilized in our work[2]: the action triples, dialogue acts, and discourse relations. The **action triples** are the “WHO-DOING-WHAT” triplets (e.g., “Sam-Asking for-Betty’s number”) in conversations that express specific socially situated identities and activities (Chen and Yang, 2021b). The **dialogue acts** describe the functions and roles of every utterance in one conversation. For example, natural conversations might often have interruption utterances with dialogue acts like *acknowledgment*, *backchannel*, *response acknowledgment* and etc. (Allen and Core, 1997; Sacks et al., 1978). **Discourse relations** describe the relations between different utterances in one conversation (Asher et al., 2016b). For example, two utterances may be related to each other with the *Question Answer Pair*.

4.2.2 Forward Process

We first add noise to prototype conversation $X_0 = \{w_0, ..., w_l\}$ to generate the noisy intermediate latent variables $X_{1:T}$ in the forward process: $X_{t+1} = q(X_t)$. To make the diffusion process more stable and efficient, the added noise needs to corrupt the prototype conversation and gives the later diffusion process appropriate flexibility to generate conversations, while avoiding removing all the prior knowledge in X_0. Thus we design and apply different types of linguistic-informed noises to perturb the structured information in conversation. Here we introduce three types of noise strategies based on the conversation structures into the forward process:

Soft-Masking Action Words For soft-masking action words, we only add noise to the action words w_i in the prototype conversation in order to perturb the action information. These action words are the words that appear in the action triples extracted from the prototype conversation using OpenIE [3] (Angeli et al., 2015; Chen and Yang, 2021b). At step t, we add a small amount of Gaussian noise to the action words w_i in the prototype conversation:

$$q_a(w_{i,t+1}|w_{i,t}) = N(w_{i,t+1}; \sqrt{(1-\beta_t)}w_{i,t}, \beta_t I) \tag{1}$$

where β_t is the amount of noise added at step t.

Soft-Masking Utterances For soft-masking utterances, we only add noise to all the words w_i in

[2] Here we choose three types of widely used structures but future work can extend this to incorporate other types of conversation structures.

[3] https://github.com/philipperemy/Stanford-OpenIE-Python

one utterance u in the prototype conversation so that the dialogue acts of the utterance are perturbed. The utterance to mask is consistent for all the steps for one prototype conversation, while we randomly reselect the utterance to mask in different epochs. At step t, we add a small amount of Gaussian noise to all the words w_i in the utterance u:

$$q_u(w_{i,t+1}|w_{i,t}) = N(w_{i,t+1}; \sqrt{(1-\beta_t)}w_{i,t}, \beta_t I) \quad (2)$$

Shuffling Discourse Relations We further randomly switch the positions of two random utterances u_i and u_j in the conversation to perturb the discourse relations in the prototype conversation. At step t, we randomly shuffle X_0:

$$q_r(X_{t+1}|X_t) = \text{Shuffle}(X_t) \quad (3)$$

In practice, we apply these three types of noises at the same time at every diffusion step t to model $q(X_{t+1}|X_t)$. Note that the forward process does not contain any trainable parameters.

4.2.3 Diffusion Process

After corrupting the hidden representations of the prototype conversation X_0 to latent variables $X_{1:T}$, we then gradually denoise X_T to $\hat{X}_0$ through diffusion steps, $\hat{X}_{t-1} = p(\hat{X}_t|\theta)$, where θ is the learned parameter to model the state transition. In practice, the transition is modeled by transformers. After every diffusion step $t \in (0, T]$, we minimize the cross entropy between the predicted conversation from $\hat{X}_{t-1}$ and the ground truth conversation C_g:

$$\mathcal{L}_t = CE(f(\hat{X}_{t-1}), C_g; \theta), t \in (0, T] \quad (4)$$

To generate desired conversation in a more controlled way, we incorporate three levels of conversation-structured information introduced in Section 4.2.1 to control the generation and we describe each of them in detail below.

Action Triples By incorporating action triples information, the conversation could include more details with diverse desired actions/events from the *token-level*. During training, we first extract such action triples $A = \{a_0, ..., a_m\}$ from the ground truth conversation C_g using OpenIE, where a_i is a "(WHO, DOING, WHAT)" triple. We then represent every triple $a_i \in A$ with the average of the output embeddings from the above $F(.)$. In order to encourage the generated conversation to describe the given actions triples, after every diffusion step $\hat{X}_{t-1} = p(\hat{X}_t|\theta), t \in (t_a, T]$, we also minimize the sum of cosine distances between the average of every token's representation in $\hat{X}_{t-1}$ and every action triple's representation:

$$\mathcal{L}_t^a = \sum_i ||\text{avg}(\hat{X}_{t-1}), F(a_i)||_{\cos}, t \in (t_a, T] \quad (5)$$

Dialogue Acts Editing the generated conversation with the desired dialogue acts information could encourage the generated conversation to be more diverse and more like human from the *utterance-level* (Allen and Core, 1997; Sacks et al., 1978). During training, we first extract the dialogue acts $D = \{d_0, ..., d_m\}$ in every ground truth conversation C_g with a learned linear dialogue acts classifier [4], where d_i is a one-hot vector that indicates the dialogue act for i-th utterance. We sum them up to represent the dialogue acts distribution in the ground truth conversation, $\hat{d} = \sum_i d_i$.

In order to encourage the generated conversation to include utterances with the given dialogue acts, we force the generated conversation to have the same dialogue acts distribution with the ground truth conversation. Specifically, after every diffusion step, $\hat{X}_{t-1} = p(\hat{X}_t|\theta), t \in (t_d, t_a]$, we first predict the dialogue acts $D^{t-1} = \{d_0^{t-1}, ..., d_n^{t-1}\}$ for every utterance in $\hat{X}_{t-1}$ with the learned classifier, where d_i^{t-1} is the predicted vector that includes the probabilities of the i-th utterance is classified as different dialogue acts. We sum the predictions $\hat{d}_{t-1} = \sum_i d_i^{t-1}$, where the j-th element in $\hat{d}_{t-1}$ denotes the total number j-type utterance in the conversation. We then minimize the L_2 distance between the ground-truth distribution and the predicted distribution from the generated conversation:

$$\mathcal{L}_t^d = ||\hat{d}, \hat{d}_{t-1}||_2, t \in (t_d, t_a] \quad (6)$$

Discourse Relations Controlling the generated conversation with the discourse relation information would encourage the utterances in it to be more related, leading to a more coherent conversation from a *conversation level*. During training, we first pre-train a discourse parsing model on a human-annotated multiparty dialogue corpus (Asher et al., 2016b) following (Shi and Huang, 2018). [5]. Via

[4] We use the hidden representations from the above $F(.)$ as inputs and we achieve the accuracy with 81.6% on Switchboard corpus, which is comparable to the state-of-the-art results (Raheja and Tetreault, 2019).

[5] We treat the hidden representations from $F(.)$ as the input. We achieve 0.781 F1 score on link predictions and 0.575 F1

Dataset	# Turns	|Conv|	|Sum|
SAMSum	10.8	129.6	23.4
DialogSum	9.8	131.0	23.6

Table 1: Data statistics of SAMSum and DialogSum including the average number of turns and words in the conversations and summaries.

this parser, we extract the discourse relation matrix $M \in R^{m\times m\times k}$ from the ground truth conversation, where m is the number of utterances and k is the total number of different discourse relations. We sum the matrix in the first two dimensions to represent the discourse relation distribution in the ground truth conversation: $\hat{r} = \sum_i \sum_j M_{i,j,k}$, where the l-th element in $\hat{r}$ means the total number of l-th discourse relation in the conversation.

We regularize the generated conversation to have the same discourse relation distribution with the ground truth conversation. After every diffusion step, $\hat{X}_{t-1} = p(\hat{X}_t|\theta), t \in (0, t_d]$, we first predict the discourse relation matrix $M_{t-1} \in R^{n\times n\times k}$ with the pre-trained parser. We also sum M_t in the first two dimensions $\hat{r}_{t-1} = \sum_i \sum_j M_{i,j,k}$ and minimize the L_2 distance between it and the ground-truth distribution:

$$\mathcal{L}_t^r = ||\hat{r}, \hat{r}_{t-1}||_2, t \in (0, t_d] \tag{7}$$

Objectives In practice, we sequentially use all three conversation structures, from lower levels to higher levels, i.e., action triples → dialogue acts → discourse relations. The order is selected through an ablation study (in Section 5.4). During training, we minimize the loss:

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_t + \sum_{t=t_a}^{T} \mathcal{L}_t^a + \sum_{t=t_d}^{t_a} \mathcal{L}_t^d + \sum_{t=1}^{t_d} \mathcal{L}_t^r \tag{8}$$

5 Experiments

5.1 Datasets and Baselines

We perform experiments on two widely-used datasets, SAMSum (Gliwa et al., 2019) and DialogSum (Chen et al., 2021), as shown in Table 1. They are originally introduced for conversation summarization, which contains open-domain real-life daily conversations with human written summaries. In this work, we reverse the datasets where we utilize the summary as input and learn the generation model to generate the long conversation. During pre-processing, we add a special token ("<s>") to indicate the begging of every utterance. We truncate the conversation into 800 tokens.

We compare our Diffuse-CG framework with several baselines:

- **BART-base** (Lewis et al., 2020): We use BART-base as our backbone model. The input only contains the summary.
- **BART-Concat**: We improve pure BART by directly concatenating controlling information including the action triples, dialogue acts and discourse relations to the end of the input summary.
- **Diffuse-CG-Con**: We use a framework similar to our Diffuse-CG while the different levels of information are combined concurrently instead of sequentially.

5.2 Experimental Setting

We initialize the prototype conversation generation model with BART-base and learn the model for 20 epochs with 3e-5 learning rate, and 0.15 warm-up ratio. The batch size is 4. For the Diffuse-CG, we utilize a 4-layer transformer whose hidden dimension is 512 to model $p(.|\theta)$. We set the diffusion steps to be $T = 500$ ($t_a = 300$ and $t_d = 100$, which means that we use 300 steps for action triples, 100 steps for dialogue acts, and 100 steps for discourse relations). We follow (Li et al., 2022b) to use an sqrt schedule in the forward process. The learning rate is set to be 3e-4 with a 0.1 warm-up ratio. The batch size is 4 and we train Diffuse-CG for 200k iterations. During inference, the beam size is set to 4. We perform all the experiments on 4 NVIDIA V100 GPUs. For diffuse-CG, the training takes around 4.8 hours, and the inference speed is 1.4second per dialogue generation.

5.3 Results

Automatic Evaluation We first evaluated all the models with:

- **ROUGE scores** (Lin and Och, 2004) measure the n-gram overlap between the generated conversation and the ground-truth conversation.
- Action coverage rate, **D**ialogue acts coverage rate, discourse **R**elation coverage measure the

score on relation classifications, which are comparable to the state-of-the-art results (Shi and Huang, 2018).

Model	Control	R-1	R-2	R-L	A Cov.	D Cov.	R Cov.	LM score	Length
BART	-	33.15	12.35	23.60	18.9	37.5	11.3	71.23	53.28
BART-Concat	a + d + r	35.32	13.38	24.75	31.5	55.5	14.9	68.14	81.36
Diffuse-CG-Con	a + d + r	38.32	17.15	26.55	33.1	72.4	23.8	69.16	83.12
Diffuse-CG †	a	38.12	18.45	27.38	**38.2**	56.3	15.1	67.76	82.42
	d	36.82	12.11	25.92	24.7	76.6	16.4	70.28	73.29
	r	37.33	15.92	24.73	20.6	68.8	27.1	69.37	78.18
	a → d	38.76	19.16	27.46	37.1	**77.9**	21.8	67.43	85.34
	a → d → r	**40.54**	**19.43**	**28.57**	36.0	75.3	**27.4**	**66.15**	**90.38**

Table 2: ROUGE-1 (↑), ROUGE-2 (↑), ROUGE-L (↑) scores, **A**ction coverage rate (↑), **D**ialogue acts coverage rate (↑), discourse **R**elation coverage rate (↑), language model scores (↓) and the length (↑) of the generated conversation for different models on the SAMSum Corpus test set. † means our model and the extra information is added in an order of action triples, dialogue acts, and discourse relations.

Model	Control	R-1	R-2	R-L	A Cov.	D Cov.	R Cov.	LM score	Length
BART	-	32.15	11.43	22.42	17.5	32.7	10.1	74.23	48.46
BART-Concat	a + d + r	32.32	14.23	23.55	30.2	51.2	16.8	70.44	83.14
Diffuse-CG-Con	a + d + r	34.52	15.18	23.22	32.0	70.8	21.6	72.16	80.34
Diffuse-CG †	a	36.56	15.45	27.38	**37.2**	58.3	15.1	71.84	82.32
	a → d	37.39	17.32	26.46	36.1	**76.9**	21.8	69.53	83.34
	a → d → r	**39.84**	**18.23**	**27.57**	35.0	75.3	**25.4**	**68.45**	**84.23**

Table 3: ROUGE-1 (↑), ROUGE-2 (↑), ROUGE-L (↑) scores, **A**ction coverage rate (↑), **D**ialogue acts coverage rate (↑), discourse **R**elation coverage rate (↑), language model scores (↓) and the length (↑) of the generated conversation for different models on the DialogSum Corpus test set.

Model	Coh.	Flu.	Fac.
BART	3.44	1.58	1.66
BART-Concat	2.89	2.17	2.14
Diffuse-CG-Con	2.43	3.88	3.52
Diffuse-CG †	**1.34**	**1.46**	**1.54**

Table 4: The average ranking every method receives in terms of **Coh**erency, **Flu**ency, **Fac**tualness from human evaluation (lower is better). † means our method.

coverage rate of the actions triples, dialogue acts, and discourse relations in the generated conversation compared to the ground-truth conversation.

- **LM score** measure the fluency by computing the perplexity from a GPT-2 pre-trained on SAMSum and DialogSum.

- **Length** measures the length of the generated conversation.

As shown in Table 2 and Table 3, we find that after adding the controlling structured information directly to the input, BART-Concat is generating better conversations compared to naive BART. This shows that our introduced conversation-structured guidance can help conversation generation by providing effective information. By applying the diffusion process, Diffuse-CG-Con and Diffuse-CG further consistently improve the performances (e.g., 8%/28%/7% improvements in ROUGE scores), which shows the effectiveness of our introduced controllable conversation generation framework. Because it makes better use of both the input summary and the controlling signals by first generating the prototype conversation and then further enriching it with the extra information using a diffusion process, which prevents the distraction from different information. Among different noise and control signals, the soft-masking action words noise and action triples diffusion worked the best, followed by shuffling discourse relations with discourse diffusion and then soft-masking utterances noise with dialogue acts diffusion. Compared to the concurrent way, our sequential Diffuse-CG works the best, indicating that editing the long conversation with a suitable order (from token levels to utterance levels and to conversation levels) is important. By gradually incorporating different levels of structure, the overall performances are improving (e.g., the ROUGE scores are increasing from 38.12/18.45/27.38 to 40.54/19.43/28.57), suggest-

Noise	R-1	R-2	R-L	A Cov.	D Cov.e	R Cov.	LM score	Length
Gaussian	33.14	14.24	23.45	34.3	70.4	23.3	75.13	75.46
Linguistic-informed †	**40.54**	**19.43**	**28.57**	**36.0**	**75.3**	**27.4**	**66.15**	**90.38**

Table 5: ROUGE-1 (↑), ROUGE-2 (↑), ROUGE-L (↑) scores, **A**ction coverage rate (↑), **D**ialogue acts coverage rate (↑), discourse **R**elation coverage rate (↑), language model scores (↓) and the length (↑) of the generated conversation for different noise strategy in Diffuse-CG on the SAMSum Corpus test set. † means our noise strategy.

Control Orders	R-1	R-2	R-L	A Cov.	D Cov.	R Cov.	LM score	Length
a → d → r †	**39.84**	**18.23**	**27.57**	**35.0**	75.3	25.4	**68.45**	**84.23**
a → r → d	37.18	16.34	24.33	34.2	**76.0**	24.3	70.13	82.48
d → a → r	35.87	14.11	25.92	33.9	73.0	24.8	71.53	82.34
d → r → a	36.84	14.24	23.38	34.1	72.6	23.1	72.45	80.58
r → a → d	37.14	15.38	25.45	33.5	75.0	**26.0**	70.18	80.38
r → d → a	38.42	16.87	26.88	31.8	74.3	25.5	69.15	78.33

Table 6: ROUGE-1 (↑), ROUGE-2 (↑), ROUGE-L (↑) scores, **A**ction coverage rate (↑), **D**ialogue acts coverage rate (↑), discourse **R**elation coverage rate (↑), language model scores (↓) and the length (↑) of the generated conversation for different orders in Diffuse-CG on the DialogSum Corpus test set. † means the best order.

ing that the sequential diffusion steps can edit the prototype conversation to higher qualities step by step, and all the introduced structures are making contributions.

Human Evaluation We conduct a human evaluation to evaluate the generated conversations qualitatively. We ask Amazon Mechanical Turk to rank the quality of 100 generated conversations (randomly sampled) from a given summary with 4 different models. Specifically, we ask them to rank them in terms of **Coherency** (the generated conversation is logical and consistent), **Fluency** (the generated conversation is reader-friendly) and **Factualness** (the generated conversation is not changing the fact from the given short descriptions). To increase annotation quality, we require turkers to have a 98% approval rate with over 10,000 approved tasks for their previous work. The pay rate was 0.5$ per hit. The rank for every summary was aggregated by majority voting. The Intra-Class Correlation (*ICC1k*) was 0.511, indicating moderate agreement (Koo and Li, 2016)). The average rank is shown in Table 4. Our Diffuse-CG achieves the best average rankings, indicating the effectiveness of incorporating conversation structures.

5.4 Ablation Study

This part describes our ablation studies on how our introduced linguistic-informed noises and the diffusion orders affect the model performances.

Noise Strategy We first visualize the performances of Diffuse-CG with different types of noise strategy in Table 5. Gaussian Noise adds Gaussian noise to all the tokens in the prototype conversation in the forward process, following previous work (Li et al., 2022b), while our introduced Linguistic-informed Noise only adds Gaussian noise to action words and random utterance as well as shuffling the conversations. Our introduced noise shows significantly better performances on SAMSum test set, indicating that our introduced noise strategy which considers the conversation structures is providing more appropriate perturbation to the prototype conversation for the diffusion process. This is because our strategy could provide flexibility to edit the prototype conversation as well as preserve the prior knowledge in the prototype conversation.

Diffusion Orders In terms of the impact of different orders to add different structured information during the diffusion process, as shown in Table 6, we find that the best overall performance is achieved by the order: action triples → dialogue acts → discourse relations, from a lower level (token/action level) to higher level (conversation-level). This might be because, in this structured order, more specific information can be introduced at the early stages when the conversations are more flexible to adopt a large amount of detailed information. When the conversation has enough information, it is then more effective to operate at a higher level like the relations between different utterances. This also indicates the effectiveness of structured ordering in general, especially when there are multiple levels of controlling information.

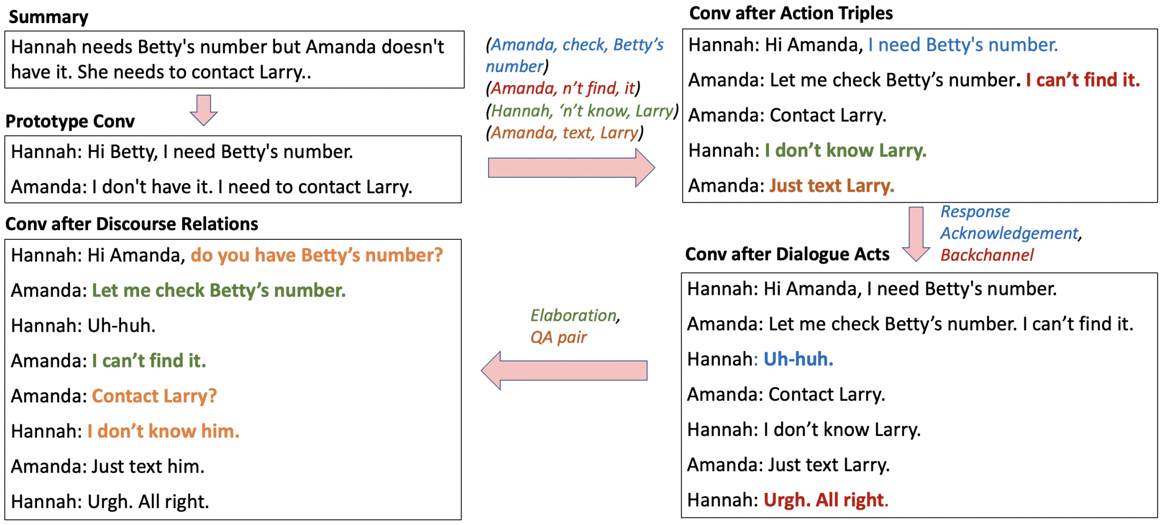
Summary
Hannah needs Betty's number but Amanda doesn't have it. She needs to contact Larry..
Prototype Conv
Hannah: Hi Betty, I need Betty's number.
Amanda: I don't have it. I need to contact Larry.
(Amanda, check, Betty's number)
(Amanda, n't find, it)
(Hannah, 'n't know, Larry)
(Amanda, text, Larry)
Conv after Action Triples
Hannah: Hi Amanda, I need Betty's number.
Amanda: Let me check Betty's number. I can't find it.
Amanda: Contact Larry.
Hannah: I don't know Larry.
Amanda: Just text Larry.
Response Acknowledgement, Backchannel
Conv after Dialogue Acts
Hannah: Hi Amanda, I need Betty's number.
Amanda: Let me check Betty's number. I can't find it.
Hannah: Uh-huh.
Amanda: Contact Larry.
Hannah: I don't know Larry.
Amanda: Just text Larry.
Hannah: Urgh. All right.
Elaboration, QA pair
Conv after Discourse Relations
Hannah: Hi Amanda, do you have Betty's number?
Amanda: Let me check Betty's number.
Hannah: Uh-huh.
Amanda: I can't find it.
Amanda: Contact Larry?
Hannah: I don't know him.
Amanda: Just text him.
Hannah: Urgh. All right.

Figure 2: An example of the generated conversation from every stage in our Diffuse-CG: the prototype conversation, the edited conversations after the action triples diffusion, the dialogue acts diffusion, and the discourse relation diffusion. The colors indicate the edits based on the provided controlling information.

5.5 Case Study

We further visualize the intermediate outputs in the diffusion process of our Diffuse-CG to interpret the generation process in Figure 2. As it shows, the prototype conversation is short and coarse. When the action information is incorporated through the first diffusion stage, the conversation is enriched by more specific action information like "*Amanda text Larry*". After the dialogue act diffusion stage, the conversation is further modified to have utterance with dialogue acts like backchannel ("*Urgh. All right*"). At last, with the discourse relation information being utilized, the conversation is more interactive and coherent with more intra-utterance relations like QA pairs. These coarse-to-fine steps show how Diffuse-CG is editing and generating better and longer conversations over time.

6 Conclusion

In this work, we introduce a novel controllable conversation generation framework that utilizes different levels of conversation structures to generate long and coherent conversations based on a given short description. Specifically, we first generate the prototype conversation and then enrich it with structure information like action triples, dialogue acts, and discourse relations, together with novel linguistic-informed noises for further adapting diffusion models to generate conversations. Experiments on SAMSum and DialogSum show the effectiveness of our framework by significantly improving over the baselines. Our proposed method also provides interpretability of how the model is gradually generating longer and better conversations.

7 Limitation

In this work, we mainly leverage control guidance such as action triples, dialogue acts, and discourse relations in structured forms that are extracted automatically from the corpus for training. We encourage future work to explore how to incorporate control information in natural language forms (for example, the natural language descriptions of the action information instead of triples). We also compose multiple modules (like the prototype generation, discourse classifier, etc.) to generate the final conversation which might lead to a larger error cascade if there is some early noise. So future work might explore how to make the pipeline learned in an end-to-end manner. What's more, we mainly focus on using three major conversation structures to help the entire conversation generation, future work might continue to explore other types of linguistic and human knowledge to further improve the conversation generation qualities.

Acknowledgements

We thank members of John Thickstun, the SALT Lab, and reviewers for their helpful feedback. This work was supported in part by an Amazon Faculty Research Award and an NSF grant IIS-2247357.

References

Tosin Adewumi, Foteini Liwicki, and Marcus Liwicki. 2022. State-of-the-art in open-domain conversational ai: A survey.

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers. Unpublished manuscript.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016a. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016b. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces.

Hengyi Cai, Hongshen Chen, Yonghao Song, Xiaofang Zhao, and Dawei Yin. 2020. Exemplar guided neural dialogue generation. In *International Joint Conference on Artificial Intelligence*.

Eugene Charniak. 1972. Toward a model of children"s story comprehension. Technical report, USA.

Jiaao Chen and Diyi Yang. 2021a. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021b. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. 2022. Continuous diffusion for categorical data.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2020. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. *arXiv preprint arXiv:2010.10044*.

James Paul Gee. 2014. *An introduction to discourse analysis: Theory and method*. Routledge.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long text generation by modeling sentence-level and discourse-level coherence.

Jian Guan, Zhenyu Yang, Rongsheng Zhang, Zhipeng Hu, and Minlie Huang. 2022. Generating coherent narratives by learning dynamic and discrete entity states with a contrastive framework.

Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Sachindra Joshi, and David Konopnicki. 2021. Summary grounded conversation generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3748–3756, Online. Association for Computational Linguistics.

Prakhar Gupta, Jeffrey Bigham, Yulia Tsvetkov, and Amy Pavel. 2021. Controlling dialogue generation with semantic exemplars. In *Proceedings of the 2021*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3018–3029, Online. Association for Computational Linguistics.

Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2022. Diffusionbert: Improving generative masked language models with diffusion models.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions.

Changzhen Ji, Yating Zhang, Xiaozhong Liu, Adam Jatowt, Changlong Sun, Conghui Zhu, and Tiejun Zhao. 2021. A neural conversation generation model via equivalent shared memory investigation. In *Proceedings of the 30th ACM International Conference on Information &amp Knowledge Management*. ACM.

Paul A Kirschner, Simon J Buckingham-Shum, and Chad S Carr. 2012. *Visualizing argumentation: Software tools for collaborative and educational sense-making*. Springer Science & Business Media.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

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

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents.

Qintong Li, Piji Li, Wei Bi, Zhaochun Ren, Yuxuan Lai, and Lingpeng Kong. 2022a. Event transition planning for open-ended text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3412–3426, Dublin, Ireland. Association for Computational Linguistics.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022b. Diffusion-lm improves controllable text generation.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics.

Phillip Lippe, Pengjie Ren, Hinda Haned, Bart Voorn, and Maarten de Rijke. 2020. Diversifying task-oriented dialogue response generation with prototype guided paraphrasing.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception.

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. 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 3727–3733, Minneapolis, Minnesota. Association for Computational Linguistics.

Ramya Ramakrishnan, Hashan Narangodage, Mauro Schilman, Kilian Weinberger, and Ryan McDonald. 2022. Long-term control for dialogue generation: Methods and evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 738–753, Seattle, United States. Association for Computational Linguistics.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Sougata Saha, Souvik Das, and Rohini Srihari. 2022. Stylistic response generation by controlling personality traits and intent. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 197–211, Dublin, Ireland. Association for Computational Linguistics.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and diverse text generation with planning-based hierarchical variational model. 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 3257–3268, Hong Kong, China. Association for Computational Linguistics.

Zhouxing Shi and Minlie Huang. 2018. A deep sequential model for discourse parsing on multi-party dialogues.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France. PMLR.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Matthew Stone, Una Stojnic, and Ernest Lepore. 2013. Situated utterances and discourse relations. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Short Papers*, pages 390–396.

Bowen Tan, Zichao Yang, Maruan AI-Shedivat, Eric P. Xing, and Zhiting Hu. 2020. Progressive generation of long text with pretrained language models.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision.

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

ACL 2023 Responsible NLP Checklist

A For every submission:

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

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

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

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

B □ Did you use or create scientific artifacts?

Not applicable. Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 4

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

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

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

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

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