A Controllable Model of Grounded Response Generation

Zeqiu Wu,[1] Michel Galley,[2] Chris Brockett,[2] Yizhe Zhang,[2] Xiang Gao,[2] Chris Quirk,[2]
Rik Koncel-Kedziorski,[1] Jianfeng Gao,[2] Hannaneh Hajishirzi,[1, 3] Mari Ostendorf,[1] Bill Dolan [2]

[1]University of Washington, Seattle, WA, USA
[2]Microsoft Research, Redmond, WA, USA
[3]Allen Institute for AI, Seattle, WA, USA
zeqiuwu1@uw.edu, mgalley@microsoft.com

Abstract

Current end-to-end neural conversation models inherently lack the flexibility to impose semantic control in the response generation process, often resulting in uninteresting responses. Attempts to boost informativeness alone come at the expense of factual accuracy, as attested by pretrained language models' propensity to "hallucinate" facts. While this may be mitigated by access to background knowledge, there is scant guarantee of relevance and informativeness in generated responses. We propose a framework that we call controllable grounded response generation (CGRG), in which lexical control phrases are either provided by a user or automatically extracted by a control phrase predictor from dialogue context and grounding knowledge. Quantitative and qualitative results show that, using this framework, a transformer based model with a novel inductive attention mechanism, trained on a conversation-like Reddit dataset, outperforms strong generation baselines.

1 Introduction

End-to-end neural models for open-domain response generation (Shang, Lu, and Li 2015; Sordoni et al. 2015; Vinyals and Le 2015; Gao, Galley, and Li 2019) are capable of generating conversational responses that are both fluent and contextually appropriate. Although the earliest neural generation models were characterized by bland and evasive responses (Li et al. 2016a), surprisingly human-like conversations can be generated using recent diversity-enhancing strategies (Holtzman et al. 2020; Gao et al. 2019a) and massive GPT-2 style models (Radford et al. 2019; Zhang et al. 2020).[1] While blandness may no longer present a challenge, the downside has been a propensity towards "hallucinated" or "fake" output (Zellers et al. 2019) of the kind illustrated in scenario I in Figure 1.

Grounded response generation (Ghazvininejad et al. 2018; Dinan et al. 2019; Qin et al. 2019) approaches can inhibit hallucination of facts. Yet grounding alone (e.g, the Wikipedia page about *La La Land* in scenario II of Figure 1) without control and semantic targeting may induce output that is accurate but vague or irrelevant. Controllable text generation (Hokamp and Liu 2017; Keskar et al. 2019; Tang et al. 2019; See et al. 2019) provides a level of semantic control that can guide the decoder towards relevant output, but in the absence of grounding the model is prevented from associating control phrases with correct facts. We posit that both grounding knowledge and lexical control are essential to generating reliable information. We therefore introduce a generation framework called controllable grounded response generation that incorporates both components. Lexical controls not only enforce response specificity, but filter lengthy, irrelevant or incoherent groundings.

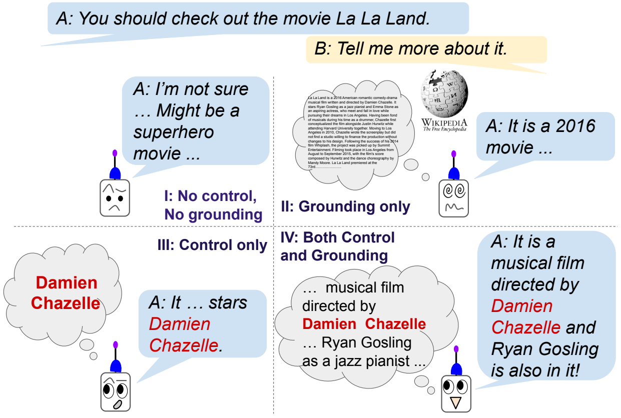
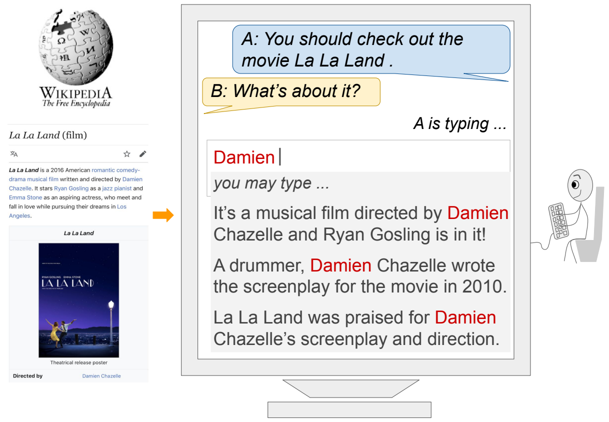

A: You should check out the movie La La Land.
B: Tell me more about it.
A: I'm not sure … Might be a superhero movie …
I: No control, No grounding
Wikipedia The Free Encyclopedia
A: It is a 2016 movie …
II: Grounding only
III: Control only
Damien Chazelle
A: It … stars Damien Chazelle.
IV: Both Control and Grounding
… musical film directed by Damien Chazelle … Ryan Gosling as a jazz pianist …
A: It is a musical film directed by Damien Chazelle and Ryan Gosling is also in it!

Figure 1: Generated responses tend to be generic or factually incorrect without grounding or control. Adding grounding improves information reliability but may lead to vague responses. Adding control boosts response specificity, but using both leads to contentful and reliable responses.

We consider two scenarios for lexical control of conversational text generation. Control can come from a human user, as in applications where an editorial assistant helps a person write a message. Figure 2 depicts a person typing keywords to indicate their semantic intent, while the machine helps construct the response to be sent out. Alternatively, control could be predicted in a fully automated system.

[1] For a related task (document creation), 72% of human judges found GPT-2 credible vs. 83% for New York Times articles: https://openai.com/blog/gpt-2-6-month-follow-up/



La La Land is a 2016 American romantic comedy-drama musical film written and directed by Damien Chazelle. It stars Ryan Gosling as a jazz pianist and Emma Stone as an aspiring actress, who meet and fall in love while pursuing their dreams in Los Angeles.

A: You should check out the movie La La Land .

B: What's about it?

Damien|

It's a musical film directed by Damien Chazelle and Ryan Gosling is in it!

A drummer, Damien Chazelle wrote the screenplay for the movie in 2010.

La La Land was praised for Damien Chazelle's screenplay and direction.

Figure 2: The machine acts as a response editorial assistant that suggests candidate responses for the user A according to the conversation history, the user's partial input (*Damien*) and grounding knowledge.

This work makes the following contributions: (1) We propose a novel framework called controllable grounded response generation (CGRG) that generates a response from the dialogue context, lexical control phrases and groundings. To the best of our knowledge, this is the first work to integrate both control and grounding into response generation, and to explore how they can be mutually beneficial. (2) We combine recent success in transformer-based generation models and a novel inductive attention mechanism to this problem setting. (3) We show through qualitative and quantitative evaluations that CGRG outperforms strong baselines where: a) the control phrases are provided by a (simulated) user, and b) automatically extracted by a control phrase prediction model.

2 Approach

We formalize the problem as follows: given dialogue context X, p lexical control phrases $C = (C_1, \cdots, C_p)$ and q sentences of grounding $G = (G_1, \cdots, G_q)$, generate a response $R = (r_1, \cdots, r_m)$ that contains semantic information guided by C. Control phrases can be either directly provided by a user or automatically derived from a control phrase predictor. The CGRG framework assumes we have a grounded conversational dataset, such as in (Qin et al. 2019). We assume that each data instance consists of a dialogue context, grounding knowledge and a reference response. To analyze this framework, we define a control mechanism that defines one or more control phrases for each instance. The controls are lexical phrases that are relevant to both the target response and some part of the grounding knowledge.

To leverage the recent success in transformer-based generation models [2] within CGRG, we concatenate X, C and G_C to be our input sequence, as shown in Figure 3 (left). Then we have the model predict the next response word given the concatenated input sequence (denoted as S) and the previous response tokens in R. G_C is the subset of G that is relevant to C. For example, in this work, we denote the grounding sentences that contain any phrase in C as G_C. To differentiate the input elements, we insert an end-of-text token $\langle eos \rangle$ at the end of each dialogue utterance in X, a $\langle c \rangle$ token at the end of each control phrase in C and a $\langle s \rangle$ token at the end of each sentence in G_C.

We first concatenate the input sequence S and the response sequence R into a long text. We denote the source sequence as $S = (w_1, \cdots, w_n)$, which is used to generate target sentence R. The conditional probability of $P(R|S)$ can be written as the product of conditional probabilities:

$$p(R|S) = \prod_{k=1}^{m+1} p(r_k | w_1, \cdots, w_n, r_1, \cdots, r_{k-1})$$

where r_{m+1} is the additional end-of-text token.

2.1 Inductive Attention

As shown in Figure 3 (left), a standard transformer-based generation model takes consecutive text sequences as input to train a language model. In our setting, we have input elements X, C, G_C in a segmented format. Simply concatenating all these input elements can induce noise, as segments may have differential relevance, and we consider attention links between such segments to be uninformative.

We remove potentially uninformative attention links for each data example by injecting pre-established structural information between C and G_C. For example, in Figure 3 (right), say that C consists of C_1, C_2, C_3, and G_C consists of G_1 and G_2. If we know C_1 is only found in G_1, then we only want to keep the attention link between C_1 and G_1, and not between C_1 and any of the other grounded sentences. Since G_C is a set of segmented sentences from G, we remove all cross-sentence links within G_C tokens. Similarly, we remove all links between non-identical phrases (e.g., tokens in C_1 do not attend to tokens in C_2). Thus, the attention links for each data example are pre-determined by structural information between C and G_C. To implement this, in each transformer layer, we manipulate attention masks by setting undesired attention links to be 0. The others remain at 1. We refer to this pre-calculated attention as inductive attention. Each response token still attends to all input tokens and other response tokens on its left.

We denote the start and end positions of a control phrase $C_i \in C$ in S by c_i^s and c_i^e, and those of a grounding sentence $G_j \in G_C$ by g_j^s and g_j^e. G_{C_i} denotes the set of grounding sentences in G_C that contain C_i. Then we calculate the attention mask M as follows:

$$M_{a,b} = \begin{cases} 0 & \text{if } a < b \\ 0 & \text{if } a \in [c_i^s, c_i^e], b \in [c_{i'}^s, c_{i'}^e], i \neq i' \\ 0 & \text{if } a \in [g_j^s, g_j^e], b \in [g_{j'}^s, g_{j'}^e], j \neq j' \\ 0 & \text{if } a \in [c_i^s, c_i^e], b \in [g_j^s, g_j^e], G_j \notin G_{C_i} \\ 1 & \text{otherwise} \end{cases}$$

Then for each transformer head, we have the stacked matrices Q, K and V to represent each example sequence (concatenated S and T) as in (Vaswani et al. 2017). We calculate

[2] Although our model can be generalized to any attention-based model, we illustrate our method with the basic auto-regressive generation model like GPT-2.

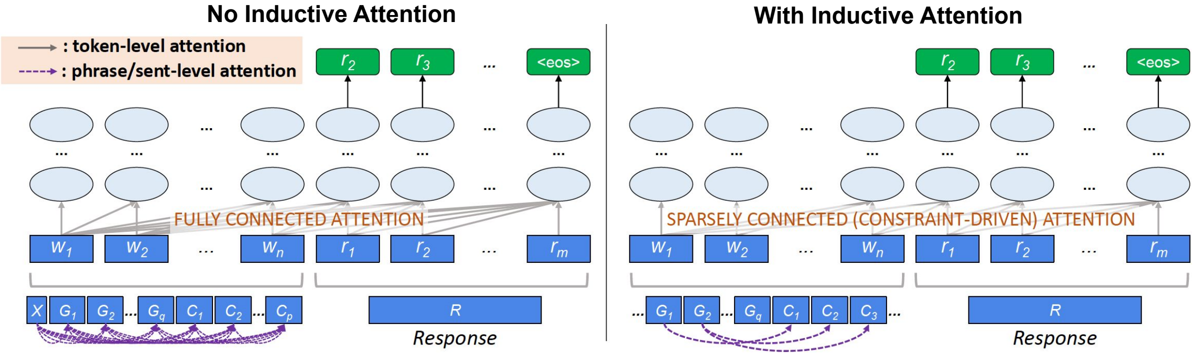

No Inductive Attention
With Inductive Attention
: token-level attention
: phrase/sent-level attention
FULLY CONNECTED ATTENTION
SPARSELY CONNECTED (CONSTRAINT-DRIVEN) ATTENTION
r_2 r_3 ... <eos>
w_1 w_2 ... w_n r_1 r_2 ... r_m
X G_1 G_2 ... G_q C_1 C_2 ... C_p
... G_1 G_2 ... G_q C_1 C_2 C_3 ...
R
Response

Figure 3: Without Inductive Attention, the model considers all possible forward attentions, which can overwhelm the model when the context contains context (X), grounding (G), and constraints (C). In contrast, Inductive Attention uses attentions that are relevant to the constraints. Each dashed arrow applies to all tokens in the corresponding X or C phrase or G grounding.

the attention as follows (d is the model dimension):

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{M \circ QK^T}{\sqrt{d}})V$$

2.2 Control Phrase Selection

User Control. Our user control is defined by lexical phrase(s). Since it is costly to have humans annotate the control phrases associated with an existing set of human comment-response pairs, we use lexical matching to simulate the human-controlled scenario. Specifically, we define control phrases C to be informative n-grams ($n \leq 5$) that appear in both the grounding document and the reference response, where informativeness is defined based on a document frequency threshold. When two n-grams are identical except for an added function word or punctuation, we use only the shorter version. In addition, we remove the matched n-grams that appear in dialogue context, with the goal of focusing on providing new information. We provide human verification details for such control phrases in Section 3.

Automatic Control Phrase Predicting. For the fully automatic scenario, we experiment with two control phrase predictors. We denote these as $\tilde{C}$ to differentiate from the user-provided C. The first predictor uses a simple retrieval-based strategy. Given a dialogue context, we rank sentences in G by IDF-weighted word overlaps with X and select the two most frequent n-grams in the top 50 sentences as $\tilde{C}$. In order to reduce the search space, we use noun phrases only.

The second predictor leverages the BERT QA system, which is fine-tuned using the dialogue context X as the query, G as the document and the control phrases in C as answers. Then we use the fine-tuned model to predict answers on test examples. We obtain the top two answers as predicted control phrases $\tilde{C}$. For both predictors, we drop the second phrase if the string fully overlaps with the first.

3 Controllable Conversation Dataset

As a social media aggregator, Reddit is effectively a dataset of multiple domains. We start with the grounded Reddit conversation dataset described in Qin et al. (2019). This dataset features Reddit conversations about web pages such as news stories and Wikipedia articles, and covers 178 subreddit topics ranging from news/technology to literature/music and multiple writing styles.

We want to focus on contexts where controllable generation is useful. Thus, we keep only responses where at least one matched phrase can be found in the grounding document. Strict lexical matching between target response and grounding assures that the retained examples have a high ratio of grounding utilization. The number of utterances of train, dev and test are 390K, 6.7K and 21K, respectively. The average length of all reference responses is 26.5, which is about 40% longer than in the full dataset due to the focus on controllability. The average numbers of phrases in C for train, dev and test set are 1.32, 1.27 and 1.38 respectively. The average numbers of sentences in G_C for train, dev and test set are 4.37, 4.32 and 4.25 respectively.

To verify that the simulated user control phrases are appropriate for the responses, we use crowd-sourced workers to annotate whether the extracted control phrases are central to the reference response, given the dialogue context. For each response, we had 3 judges to enter a score on a scale of 1 (completely unrelated) to 6 (very central), where 5 was "somewhat central" and 4 was "neutral." In 2000 annotated examples, the median score was 4.33 and 67.4% of examples had a score over 4. Inter-rater agreement was "fair" with Krippendorff's alpha coefficient at 0.32.[3]

4 Experimental Setup

4.1 Training and Inference Setup

In our experiments, all transformer-based models have both type and positional embedding for each input token. We treat X, each sentence in G_C, each phrase in C and response R as separate segments. We set the maximum number of sentences in G_C to be 20 and maximum number of phrases in C

[3] This dataset is a filtered version of (Qin et al. 2019)'s public dataset. To further facilitate reproducibility, we release our data preparation and modeling code at https://github.com/ellenmellon/CGRG.

to be 10, then we have "0" for X; "1-20" for G_C; "21-30" for C and "31" for R tokens as type embedding. For each segment, we have the position embedding for each token as its position in that segment.

We use GPT-2 as the basic transformer-based generation model in our experiments for drawing fair comparison with the DialoGPT (Zhang et al. 2020) architecture, a state-of-the-art conversational response generation model trained on 147M Reddit comment chains on the basis of GPT-2. We use the small version of GPT-2 with 117M parameters, with the maximum length of the input or target response sequence to be 512. We use BPE tokenization, following GPT-2. We finetune our model and all other GPT-2-based baselines (including DialoGPT) on our controllable and grounded Reddit dataset on top of the original DialoGPT. None of their Reddit training or validation examples overlap with our test examples. We use batch size 32. Learning rate (1e-5) and warmup steps (1600) are tuned on the dev set perplexity, with all other parameters being the same as DialoGPT [4]. Each training process is run on 2 Tesla K-80 nodes.

We use greedy search as the decoding strategy for all GPT-2 and GPT2IA setups, except for a single experiment setting where grid beam search (GBS) (Hokamp and Liu 2017) is applied for comparison with lexically constrained decoding. We also compare our methods with GBS to investigate whether it helps to encode the constraints into the hidden state during both training and inference, as GBS uses lexical constraints only during inference.

4.2 Evaluated Systems

Models Although our designed model can be applied to any transformer-based generation model, we use GPT-2 as our base model for experiments. We use the following models for experiments: (a) **GPT-2**, which has the same architecture as DialoGPT (Zhang et al. 2020), under the input setting X (see below) or the first line in both Table 1 and Table 3; (b) **GPT2IA** model with inductive attention; (c) **GPT-2 + GBS** that applies the attended GPT-2 model, while control phrases C are given to the model at decoding time;[5] and (d) **CMR** (Qin et al. 2019), which is a previous state-of-the-art grounded response generation model on the Reddit dataset that combines a MRC model and an LSTM decoder.

Input Settings We evaluate the above models according to the following settings to analyze how control and grounding help improve the response generation performance:

- X**:** This is the standard setting for non-controllable response generation, where only the dialogue context is given. We conduct experiments for the GPT-2 generation model. Note that GPT-2 in this setting is the same as the DialoGPT architecture.
- X+G**:** This is the standard setting for grounded response generation. We compare two models: CMR and GPT-2. GPT-2 for this setting concatenates X and G as its input. As both models have an input sequence length limit, only a random subset of grounding is fed into each model.
- X+C**:** This is the controllable response generation setting without grounding. We conduct GPT-2 experiments by concatenating X and C.
- X+G_C**:** This setting measures how the grounding only relevant to C can help with response generation, without explicitly providing C. We conduct experiments for GPT-2, by concatenating X and G_C as the input.
- X+C+G_C**:** This setting measures how grounded control can help with response generation. We conduct experiments for GPT-2 and GPT2IA, by concatenating X, G_C and C as the input.
- X+C+G**:** This setting is for comparison against existing constrained generation methods like grid beam search (GBS) introduced in Hokamp and Liu (2017), where lexical control phrases are added in decoding only without involving training. We conduct experiments for GPT-2 where X and G are the only encoded inputs and C is only applied in decoding with GBS.

4.3 Automatic Evaluation

Previous work (Li et al. 2016b; Sun and Nenkova 2019) has shown that automatic metrics for generation can be unreliable and have low absolute values, so we rely on human evaluation for our main conclusions. However, due to the high cost of human evaluation, automatic evaluation metrics can be useful for hyper-parameter tuning and model selection. For automatic evaluation, we measure the relevance of the generated responses with three metrics: BLEU-4 (Papineni et al. 2002), NIST-4 (Doddington 2002) (a variant of BLEU that weights n-gram matches by their information gain, penalizing uninformative n-grams), and diversity of bigrams in generated responses (Div-2, the ratio between the number of distinct vs. total bigrams). The human evaluation is used to verify the improvements of the best case systems as determined by the automatic metrics.

We experiment with both user-controllable and fully automatic response generation, with simulated user-selected and predicted lexical control phrases, respectively. As different reference responses correspond to different "gold" control phrases, we use single-reference evaluation for the user-controllable setting. Predicted control phrases are independent of reference responses, so we use multi-reference evaluation. Comments in Reddit discussions are often associated with multiple responses, which provide a multi-reference test set. For each metric, we report the highest score among up to 5 alternative human references.

4.4 Human Evaluation

Human evaluation was conducted using crowd-sourced workers. Judges were presented with paired randomized outputs. The document title, a short snippet of the document and up to two conversational turns were provided as context. Relevance and appropriateness to the preceding dialog

[4] All model configurations in Section 4.2 share the same hyper-parameter values (only tuned 3 times on top of the original values from DialoGPT) except CMR, which has the same parameters in the original paper.

[5] Such constrained decoding is based on grid beam search (GBS) introduced in Hokamp and Liu (2017), where lexical control phrases are added in decoding only, without involving training.

Setting	Model	NIST	BLEU	Div-2	Avg-L
1) X	GPT-2	0.90	0.55%	4.9%	22.2
2) X+G	CMR	0.34	0.17%	11.3%	15.1
3) X+G	GPT-2	0.98	0.67%	7.5%	23.1
4) X+C	GPT-2	1.67	2.65%	10.7%	28.7
5) X+G_C	GPT-2	1.34	1.58%	11.1%	26.6
6) X+C+G	GPT-2+GBS[7]	1.60	2.38%	10.6%	26.8
7) X+C+G_C	GPT-2	1.77	3.22%	11.3%	27.0
8) X+C+G_C	GPT2IA	**1.80**	**3.26**%	**11.6**%	25.9

Table 1: User-controllable Response Generation automatic evaluation.

	GPT2IA	Tied	GPT-2	
Relevance: *Which response is more relevant and appropriate to the preceding dialog?*				
X+C+G_C	**69.8**%	14.1%	16.1%	X+C+G+GBS
X+C+G_C	**42.1**%	23.5%	34.4%	X+C
X+C+G_C	**38.1**%	28.6%	33.3%	X+C+G_C
Consistency: *Which response is more consistent with the grounding text?*				
X+C+G_C	28.1%	44.3%	27.6%	X+C+G_C
X+C+G_C	**37.6**%	31.4%	31.0%	X+C

Table 2: Controllable Response Generation human evaluation for relevance and background consistency, showing preferences (%). A number in bold indicates that the system is significantly better at $p \leq 10^{-5}$, computed using 10k bootstrap replications.

and consistency with the background text (as a metric of factual correctness) were measured. Judgments were based on a five-point Likert scale, and ties were permitted. Three to four judges evaluated each pair, and metrics were imposed to block poorly performing judges. Inter-rater agreement was "fair" with Krippendorff's coefficient at 0.32.[6]

5 Results and Analysis

5.1 User-controlled Response Generation

The user-controllable grounded response generation experiments are summarized in Table 1, using single-reference evaluation. The low BLEU/NIST scores are consistent with differences between human references as seen in Sec. 5.2.

Lines 1-3 are not controllable settings, while lines 4-8 have control phrases as input, either explicitly or implicitly. The performance gap between lines (1-3) and (4-8) demonstrates the value of adding control. Additionally, we can draw the following conclusions by comparing rows in Table 1: (i) **1 vs. 3**: Simply adding grounding to the model input improves the performance somewhat; (ii) **2 vs. 3**: GPT-2 in general performs better than the LSTM-based model CMR, indicating that the combination of pre-training and having a transformer-based decoder helps improve generation; (iii) **3 vs. 5**: providing constraint-sensitive grounding boosts performance compared to having all the grounding (iv) **5 vs. 7-8**: providing control phrases in an explicit way is important; (v) **6 vs. 7-8**: applying control in hidden states helps the model generate better quality responses than applying control at decoding only; (vi) **7 vs. 8**: inductive attention helps reduce noise and improve the performance.

Setting	Model	Phrase Predictor	NIST	BLEU	Div-2
X	GPT-2	-	1.42	**1.31**%	18.1%
X+$G_{\tilde{C}}$	GPT-2	Retrieval-based	1.61	1.26%	19.4%
X+$\tilde{C}$+$G_{\tilde{C}}$	GPT2IA	Retrieval-based	**1.67**	1.23%	**20.2**%
X+$\tilde{C}$+$G_{\tilde{C}}$	GPT2IA	BertQA	**1.67**	1.26%	19.6%
Human	-	-	2.04	2.56%	62.8%

Table 3: Response Generation automatic evaluation (multi-references) using constraints from control phrase predictor. Note that results of Tables 1 and 3, as user constraints give away significant information about the intended response.

Comparing lines **6 vs. 7-8** we see that applying control in hidden states is more effective than strict constraints at decoding, but it is possible that controls at the training and decoding stages could be complementary. Investigation of methods of combining these are left to future research.

Human evaluation results in Table 2 confirm that X+C+G_C+GPT2IA outperforms other systems, except in the case of Consistency, where there is no statistical difference between X+C+G_C+GPT2IA and X+C+G_C+GPT2, both grounded systems.

5.2 Predictor-controlled Response Generation

In the fully automatic response generation scenario, we compare two models for predicting control phrases. Table 3 compares the two models to the setting where no control phrases are provided to the model, using multi-reference evaluation. We observe that both the retrieval-based and BERT QA based control phrase predictors outperform X+GPT-2 (DialoGPT) and achieve good Div-2 results, with NIST scores similar to the user-controlled settings but low BLEU score. For more insight into automatic scores, we also report scores on human responses. When defining the multi-reference responses, we hold out one response in each set as the "human" system setting. These human responses have higher diversity, but their NIST and BLEU scores remain low owing to the huge range of possible responses to any comment.

In paired comparisons using human judges our fully automatic system X+$\tilde{C}$+$G_{\tilde{C}}$ +GPT2IA was rated as having the most informative and relevant response in 32% of cases, significantly better than the 20-21% for X+GPT-2 (DialoGPT).

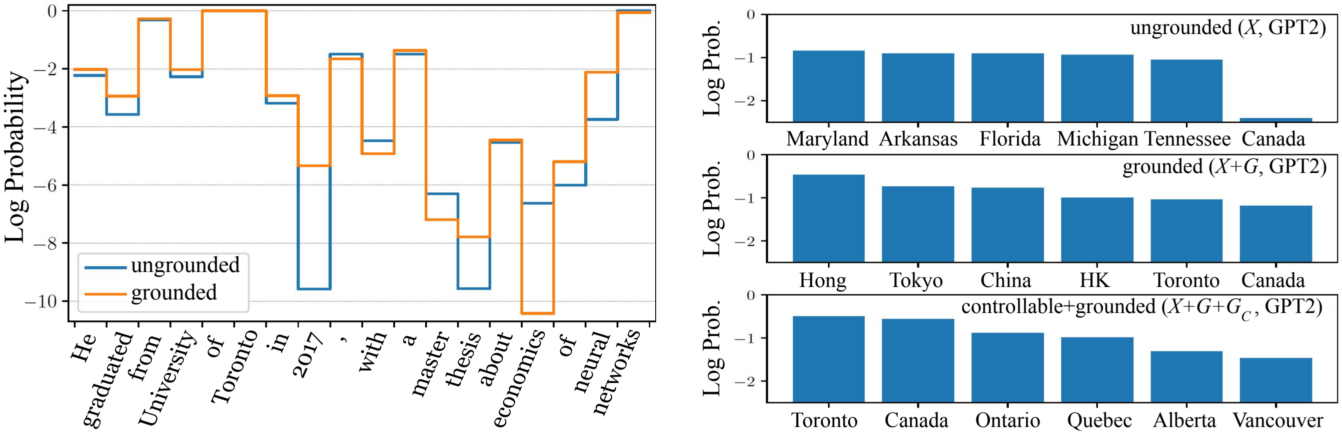
5.3 Qualitative Analysis

To understand how grounding knowledge assists generation, we plot the token-level probability (Figure 4a) for both X+C

[6] Sample sizes vary. The number was reduced from an initial 1,000 when we automatically removed a number of instances where egregiously offensive content rendered them inappropriate to display to judges.

[7] X+C+G (GBS) only takes X+G as the encoder input while C is seen at decoding only.

Grounding (G_C)**:** Sam got his bachelor degree in Physics at University of Science and Technology of China. He spent 6 months at University of Tokyo in Japan as a visiting student, when he was a master student in Computer Science at University of Hong Kong from 2010-2012. And he finished his PhD at University of Toronto in Canada with his research focused on interpretability of neural networks on text generation in 2017.
Context (X)**:** Do you know the education background of the new faculty, Sam?
Control phrases (C)**:** University of Toronto; Neural networks
Model predictions:



(a) The grounded model (X+C+G_C+GPT2IA) offers better discrimination vis-à-vis an ungrounded model (X+C+GPT2), given a document about a person's education background.

(b) The top 5 tokens (plus *Canada*) generated after the partial response *Sam just graduated from University of.* The ungrounded model prefers generic predictions. The grounded model is more topically relevant. The constraint further positively influences the hidden state.

Figure 4: Effect of grounding and control on text generation.

and X+C+G_C systems. We intentionally select an example about an uncommon entity to eliminate the possibility that the knowledge is captured in pre-training. The figure shows the token-level probability of a potential response, given a dialogue context, two control phrases, and grounding sentences. The grounded model assigns higher probabilities to contextual words from grounding such as *graduated* and *thesis* and to factually correct entity tokens like *2017*. It assigns lower probability to factually incorrect tokens like *economics*. These observations suggest that grounding knowledge can help controllable generation: contextualize control phrases and distinguish correct vs. incorrect facts.

Figure 4b illustrates the functions of control and grounding. We list the top 6 tokens after a partial response given the same dialogue context and grounding, and control phrase *Canada*. The ungrounded non-controllable model gives equally distributed probabilities to well-known American state names after *University of.* Adding grounding helps the model rank locations based on background knowledge. Further adding controls helps the model locate the correct or intended answer.

To quantify the observations in Figure 4a and Figure 4b, we sample 100 test examples and randomly pick an entity from each reference response to calculate the entity's probability from each model. We restrict the entity to be non-occurring in control phrases. Then we calculate the average probability ratio for X+C/X+C+G_C and X+G/X+C+G_C, to be 0.773 and 0.886 respectively. Both of them are smaller than 1.0, which indicates having both grounding and control phrases gives higher probability to correct entities than either of these alone. Explicit control phrases can be leveraged to dissect the generation process. Table 4 shows how controls may guide or perturb the GPT2IA model to produce responses with diverging semantics. An example with sample outputs of different systems in the user-controlled scenario is shown in Table 5.

6 Related Work

6.1 Grounded Response Generation

Although some relevant work draws on external knowledge sources, none incorporates user control. Ghazvininejad et al. (2018) develop a memory network based model that leverages grounding information from Foursquare tips. Moghe et al. (2018) and Zhou, Prabhumoye, and Black (2018) collect movie discussion datasets via crowdsourcing. These are limited to specific domains. Dinan et al. (2019) crowdsource conversations where each utterance is grounded in up to one single sentence. We focus on a more realistic, scalable setting in which a response may constitute a blend of multiple grounding informations, rather than a single factual sentence rephrasing. Other researchers propose a copy mechanism to import tokens from dialogue context and grounding (Yavuz et al. 2019) or leverage a reading comprehension model to co-encode dialogue context and grounding (Qin et al. 2019).

Other work incorporates relational knowledge bases (Zhu

Dialogue Context	With “nihonium”, Japanese scientists become first from an Asian country to name atomic element.
Control	**periodic table**
Grounding	... The **periodic table** is a great legacy in chemistry ...
X+C+G_C +GPT2IA	I’m not sure if this is a good thing or not, but I’m pretty sure the **periodic table** is a great legacy in chemistry.
Control	**artificially**
Grounding	... The **artificially** synthesized element has 113 protons in its nucleus ...
X+C+G_C +GPT2IA	I wonder if they will be able to name a chemical that **artificially** produces atomic elements.

Table 4: For the same dialogue context, GPT2IA generates varied responses given different control phrases.

Context	76 % of all known serial killers in the 20th century were from the United States.
Control	**law enforcement**
Grounding	... and may include more stringent policies for military personnel in **law enforcement** or security ... Should the cases cross multiple jurisdictions, the **law enforcement** system in **the US** is fragmented and thus not configured to detect multiple similar murders across a large geographic area ...
X	I’m pretty sure **the US** had a police force of around 100,000 people.
XC	I’m pretty sure the USA had a large number of serial killers in **the US**. I’m sure the USA had a large number of **law enforcement** officers in **the US**.
XCG_C	I’m not sure if this is true, but I’m pretty sure that **the US** has a lot of **law enforcement** officers that are from **the US**.
XCG_C +IA	I’m not sure if this is true, but I’m pretty sure that the **law enforcement** in **the US** is not very good at **detecting** serial killers.

Table 5: Sample outputs of the systems, with baseline outputs for comparison.

et al. 2017; Liu et al. 2018) or commonsense knowledge graphs (Young et al. 2018) to conversational models. More recently, Liu et al. (2019) develop a graph-path-based method on knowledge graphs augmented with unstructured grounding. Our present work focuses on text based grounding and does not require preconstructed knowledge graphs.

6.2 Controlled and Content-Planned Generation

Prior work on machine translation and language generation has sought to enforce user-specified constraints, primarily in the form of lexical constraints (Hokamp and Liu 2017; Hu et al. 2019b,a; Miao et al. 2019). These approaches exploit constraints at inference time only; in our case, constraints are applied during training, with the option also of application at inference. Application during training enables the constraints to be incorporated into the latent space for better predictions.

In other related work, (See et al. 2019; Keskar et al. 2019; Tang et al. 2019) have explored non-lexical constraints, but do not examine how these could facilitate use of grounding and external knowledge. We see this line of research as complementary to ours. These papers also make the assumption that (gold) constraints can always given to the system, which limits the potential to demonstrate broader benefits of the approaches. To address this concern, we also evaluate our models in settings where *gold* constraints are unavailable (e.g., based on predicted constraints produced by a control phrase predictor).

Controllable text generation has also been employed in text style transfer (Hu et al. 2017) and other tasks (Ficler and Goldberg 2017; Dong et al. 2017; Gao et al. 2019b), to disentangle high-level style information from contextual information such that the former can be independently manipulated. (Zhao, Lee, and Eskenazi 2018) uses discrete latent actions to learn an interpretable representation for task-oriented dialogue systems. While these works use “style” labels (e.g. positive/negative, formal/informal) as controlling signals, our framework controls generation with specific lexical constraints, allowing for fine-grained semantic control.

Content planned generation (Hua and Wang 2019; Wiseman, Shieber, and Rush 2017) targets selection of a few keyphrases or table entries as the focus of text generation. However this line of work does not need to consider dialogue context, which is essential for response generation.

7 Conclusion

The CGRG framework allows users to inject soft semantic control into the generation process. It incorporates grounding to contextualize users’ semantic intents as well as to boost information reliability. We introduce an inductive attention mechanism for self-attention-based generation models like GPT-2 in order to boost its performance. We also demonstrate that this framework can benefit standard automatic response generation when integrated with a control phrase predictor. Some interesting future directions include exploring various types of user desired control and extending the controllable grounded generation concept to broader generation tasks like document writing assistance.

Acknowledgements

We thank members of Microsoft Research and University of Washington's NLP groups who provided feedback and insights to this work.

References

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational agents. In *ICLR*.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proc. of HLT*, 138–145.

Dong, L.; Huang, S.; Wei, F.; Lapata, M.; Zhou, M.; and Xu, K. 2017. Learning to generate product reviews from attributes. In *Proc. of EACL*, 623–632.

Ficler, J.; and Goldberg, Y. 2017. Controlling Linguistic Style Aspects in Neural Language Generation. In *Proc. of EMNLP*, 94–104.

Gao, J.; Galley, M.; and Li, L. 2019. Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval* 13(2-3): 127–298.

Gao, X.; Lee, S.; Zhang, Y.; Brockett, C.; Galley, M.; Gao, J.; and Dolan, B. 2019a. Jointly Optimizing Diversity and Relevance in Neural Response Generation. In *Proc. of NAACL*, 1229–1238.

Gao, X.; Zhang, Y.; Lee, S.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2019b. Structuring Latent Spaces for Stylized Response Generation. In *Proc. of EMNLP*, 1814–1823.

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; tau Yih, W.; and Galley, M. 2018. A Knowledge-Grounded Neural Conversation Model. In *Proc. of AAAI*, 5110–5117.

Hokamp, C.; and Liu, Q. 2017. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. In *Proc. of ACL*, 1535–1546.

Holtzman, A.; Buys, J.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *ICLR*.

Hu, J. E.; Khayrallah, H.; Culkin, R.; Xia, P.; Chen, T.; Post, M.; and Van Durme, B. 2019a. Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting. In *Proc. of NAACL*, 839–850.

Hu, J. E.; Rudinger, R.; Post, M.; and Durme, B. V. 2019b. ParaBank: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation. In *Proc. of AAAI*, 6521–6528.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward Controlled Generation of Text. In *Proc. of ICML*, 1587–1596.

Hua, X.; and Wang, L. 2019. Sentence-Level Content Planning and Style Specification for Neural Text Generation. In *Proc. of EMNLP*, 591–602.

Keskar, N. S.; McCann, B.; Varshney, L. R.; Xiong, C.; and Socher, R. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *Computing Research Repository* arXiv:1909.05858. Version 2.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proc. of NAACL*, 110–119.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016b. A persona-based neural conversation model. In *Proc. of ACL*, 994–1003.

Liu, S.; Chen, H.; Ren, Z.; Feng, Y.; Liu, Q.; and Yin, D. 2018. Knowledge Diffusion for Neural Dialogue Generation. In *Proc. of ACL*, 1489–1498.

Liu, Z.; Niu, Z.-Y.; Wu, H.; and Wang, H. 2019. Knowledge Aware Conversation Generation with Explainable Reasoning over Augmented Graphs. In *Proc. of EMNLP*, 1782–1792.

Miao, N.; Zhou, H.; Mou, L.; Yan, R.; and Li, L. 2019. CGMH: Constrained Sentence Generation by Metropolis-Hastings Sampling. In *Proc. of AAAI*, 6834–6842.

Moghe, N.; Arora, S.; Banerjee, S.; and Khapra, M. M. 2018. Towards Exploiting Background Knowledge for Building Conversation Systems. In *Proc. of EMNLP*, 2322–2332.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*, 311–318.

Qin, L.; Galley, M.; Brockett, C.; Liu, X.; Gao, X.; Dolan, B.; Choi, Y.; and Gao, J. 2019. Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading. In *Proc. of ACL*, 5427–5436.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*. Accessed 22 March 2021.

See, A.; Roller, S.; Kiela, D.; and Weston, J. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proc. of NAACL*, 1702–1723.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural Responding Machine for Short-Text Conversation. In *Proc. of ACL-IJCNLP*, 1577–1586.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proc. of NAACL-HLT*, 196–205.

Sun, S.; and Nenkova, A. 2019. The Feasibility of Embedding Based Automatic Evaluation for Single Document Summarization. In *Proc. of EMNLP*, 1216–1221.

Tang, J.; Zhao, T.; Xiong, C.; Liang, X.; Xing, E. P.; and Hu, Z. 2019. Target-Guided Open-Domain Conversation. In *Proc. of ACL*, 5624–5634.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, 5998–6008.

Vinyals, O.; and Le, Q. 2015. A Neural Conversational Model. In *Proc. of ICML Deep Learning Workshop*.

Wiseman, S.; Shieber, S.; and Rush, A. 2017. Challenges in data-to-document generation. In *Proc. of EMNLP*, 2253–2263.

Yavuz, S.; Rastogi, A.; Chao, G.-L.; and Hakkani-Tur, D. 2019. DeepCopy: Grounded Response Generation with Hierarchical Pointer Networks. In *Proc. of SIGdial Meeting on Discourse and Dialogue*, 122–132.

Young, T.; Cambria, E.; Chaturvedi, I.; Huang, M.; Zhou, H.; and Biswas, S. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proc. of AAAI*, 4970–4977.

Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019. Defending Against Neural Fake News. In *NeurIPS*, 9051–9062.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proc. of ACL System Demonstrations*, 270–278.

Zhao, T.; Lee, K.; and Eskenazi, M. 2018. Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation. In *Proc. of ACL*, 1098–1107.

Zhou, K.; Prabhumoye, S.; and Black, A. W. 2018. A Dataset for Document Grounded Conversations. In *Proc. of EMNLP*, 708–713.

Zhu, W.; Mo, K.; Zhang, Y.; Zhu, Z.; Peng, X.; and Yang, Q. 2017. Flexible End-to-End Dialogue System for Knowledge Grounded Conversation. *Computing Research Repository* arXiv:1709.04264.