Thinking Clearly, Talking Fast: Concept-Guided Non-Autoregressive Generation for Open-Domain Dialogue Systems

Yicheng Zou, Zhihua Liu, Xingwu Hu, Qi Zhang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University
Shanghai, China
{yczou18,liuzh20,xwhu20,qz}@fudan.edu.cn

Abstract

Human dialogue contains evolving concepts, and speakers naturally associate multiple concepts to compose a response. However, current dialogue models with the seq2seq framework lack the ability to effectively manage concept transitions and can hardly introduce multiple concepts to responses in a sequential decoding manner. To facilitate a controllable and coherent dialogue, in this work, we devise a concept-guided non-autoregressive model (CG-nAR) for open-domain dialogue generation. The proposed model comprises a multi-concept planning module that learns to identify multiple associated concepts from a concept graph and a customized Insertion Transformer that performs concept-guided non-autoregressive generation to complete a response. The experimental results on two public datasets show that CG-nAR can produce diverse and coherent responses, outperforming state-of-the-art baselines in both automatic and human evaluations with substantially faster inference speed.

1 Introduction

Creating a "human-like" dialogue system is one of the important goals of artificial intelligence. Recently, due to the rapid advancements in natural language generation (NLG) techniques, data-driven approaches have attracted lots of research interest and have achieved impressive progress in producing fluent dialogue responses (Shang et al., 2015; Vinyals and Le, 2015; Serban et al., 2016; Li et al., 2016). However, such seq2seq models tend to degenerate generic or off-topic responses (Tang et al., 2019; Welleck et al., 2020). An effective way to address this issue is to leverage external knowledge (Zhou et al., 2018a,b) or topic information (Xing et al., 2017), which are integrated as additional semantic representations to improve dialogue informativeness.

Although promising results have been obtained by equipping dialogue models with external knowledge, the development of dialogue discourse still has its own challenge: human dialogue generally evolves around a number of concepts that might frequently shift in a dialogue flow (Zhang et al., 2020). The lack of concept management strategies might lead to incoherent dialogue due to the loosely connected concepts. To address this problem, recent studies have combined concept planning with response generation to form a more coherent and controllable dialogue (Wu et al., 2019; Xu et al., 2020a,b; Wu et al., 2020; Zhang et al., 2020).

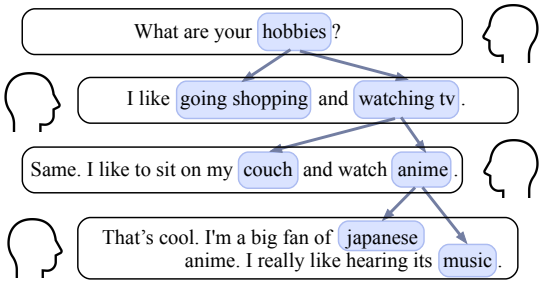


Figure 1: An exemplar dialogue with concept transitions, where each utterance is composed of multiple associated concepts to convey diverse information.

Most of these approaches incorporate concepts into responses in an implicit manner, which cannot guarantee the appearance of a concept in a response. Compared with dialogue concepts, a large proportion of chit-chat words are common and usually have a high word frequency and are relatively over-optimized in language models (Gong et al., 2018; Khassanov et al., 2019). Consequently, conventional seq2seq generators are more "familiar" with these generic words than those requiring concept management, which prevents introducing certain concepts to the response with sequential decoding (either greedily or with beam search) (Mou et al., 2016). Moreover, speakers naturally associate multiple concepts to proactively convey diverse information, e.g., action, entity, and emotion (see Figure 1). Unfortunately, most existing methods

can only retrieve one concept for each utterance (Tang et al., 2019; Qin et al., 2020). Another line of approaches attempt to explicitly integrate concepts into responses and generate the remaining words in both directions (Mou et al., 2016; Xu et al., 2020a), but they also fail to deal with multiple concepts.

In this paper, we devise a concept-guided non-autoregressive model (CG-nAR) to facilitate dialogue coherence by explicitly introducing multiple concepts into dialogue responses. Specifically, following Xu et al. (2020a), a concept graph is constructed based on the dialogue data, where the vertices represent concepts, and edges represent concept transitions between utterances. Based on the concept graph, we introduce a novel multi-concept planning module that learns to manage concept transitions in a dialogue flow. It recurrently reads historical concepts and dialogue context to attentively select multiple concepts in the proper order, which reflects the transition and arrangement of target concepts. Then, we customize an Insertion Transformer (Stern et al., 2019) by initializing the selected concepts as a partial response for subsequent non-autoregressive generation. The remaining words of a response are generated in parallel, aiming to foster a fast and controllable decoding process.

We conducted experiments on Persona-Chat (Zhang et al., 2018) and Weibo (Shang et al., 2015). The results of automatic and human evaluations show that CG-nAR achieves better performance in terms of response diversity and dialogue coherence. We also show that the inference time of our model is much faster than conventional seq2seq models. All our codes and datasets are publicly available.[1]

Our contributions to the field are three-fold: 1) We design a concept-guided non-autoregressive strategy that can successfully integrate multiple concepts into responses for a controllable decoding process. 2) The proposed multi-concept planning module effectively manages multi-concept transitions and remedies the problem of dialogue incoherence. 3) Comprehensive studies on two datasets show the effectiveness of our method in terms of response quality and decoding efficiency.

2 Related Work

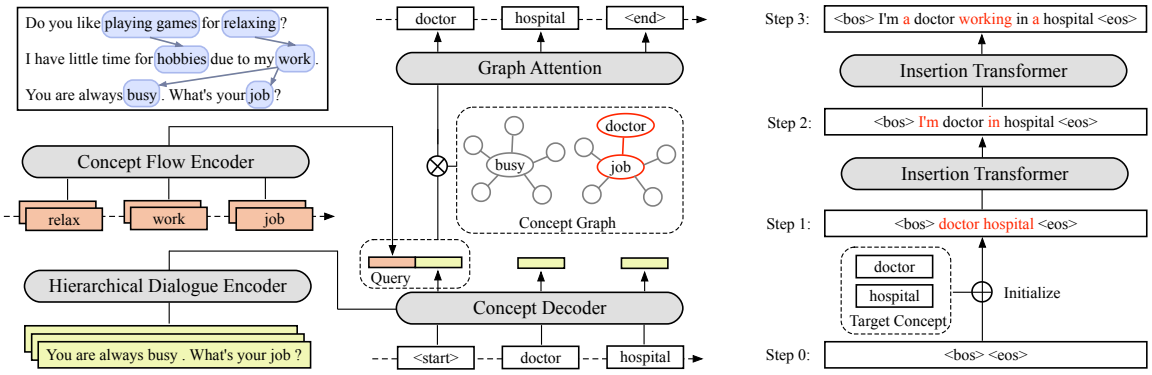
2.1 Open-Domain Dialogue Generation

Neural seq2seq models (Sutskever et al., 2014) have achieved remarkable success in dialogue systems (Shang et al., 2015; Vinyals and Le, 2015; Serban et al., 2016; Xing et al., 2017), but they prefer to produce generic and off-topic responses (Tang et al., 2019; Welleck et al., 2020). Dozens of works have attempted to incorporate external knowledge into dialogue systems to improve informativeness and diversity (Zhou et al., 2018a; Zhang et al., 2018; Dinan et al., 2019; Ren et al., 2020). Beyond the progress on response quality, a couple of works focus on goal planning or concept transition for a controllable and coherent dialogue (Yao et al., 2018; Moon et al., 2019; Wu et al., 2019; Xu et al., 2020a,b; Wu et al., 2020; Zhang et al., 2020). Most of these works mainly explore how to effectively leverage external knowledge graphs and extract concepts from them. Nevertheless, they generally introduce concepts into the response implicitly with gated controlling or copy mechanism, which cannot ensure the success of concept integration because seq2seq models prefer generic words. Some works (Mou et al., 2016; Xu et al., 2020a) try to produce concept words first and generate the remaining words to both directions to complete a response, but they cannot handle the situation of multiple concepts. By contrast, we focus on how to effectively integrate multiple extracted concepts into dialogue responses. The proposed CG-nAR applies the non-autoregressive mechanism, which can explicitly introduce multiple concepts simultaneously to responses to enhance coherence and diversity.

2.2 Non-Autoregressive Generation

Compared with traditional sequential generators that conditions each output word on previously generated outputs, non-autoregressive (non-AR) generation avoids this property to speed up decoding efficiency and has recently attracted much attention (Gu et al., 2018, 2019; Ma et al., 2019; Stern et al., 2019). Another relevant line of research is refinement-based generation (Lee et al., 2018; Kasai et al., 2020; Hua and Wang, 2020; Tan et al., 2021), which gradually improves generation quality by iterative refinement on the draft instead of one-pass generation. For dialogue systems, there has been prior works that attempt to improve the traditional autoregressive generation. Mou et al. (2016) explores the way of generating words to both directions, but it is still in an autoregressive manner. Song et al. (2020) introduces a three-stage refinement strategy for improving persona

[1] https://github.com/RowitZou/CG-nAR



(a) Multi-Concept Planning

(b) Concept-Guided Non-AR Generation

Figure 2: The overall framework of CG-nAR. (a) The multi-concept planning module conditions on the previous concept flow and the dialogue context to attentively select multiple associated concepts from the concept graph. (b) The selected concepts are used to initialize a partial response for subsequent non-autoregressive generation.

consistency of dialogue generation, but it requires a specialized consistency matching model for inference. Han et al. (2020) applies the non-AR mechanism to dialogue generation, aiming to alleviate the non-globally-optimal issue to produce a more diverse response. In this work, we further use dialogue concepts to guide response generation. We customize an Insertion Transformer (Stern et al., 2019) and arrange dialogue concepts as a partial input sequence, which is different from the original setting where texts are generated from scratch. By this means, multiple concepts can be naturally introduced to the response as a guidance to foster a more controllable non-AR generation.

3 Methodology

The overall framework of CG-nAR is shown in Figure 2. Based on a concept graph that represents candidate concept transitions, a multi-concept planning module is designed to select and arrange appropriate target concepts from the contextually related subgraphs, which is conditioned on the previous concept flow and the dialogue context. Then, we input the selected concepts as a partial response into an Insertion Transformer (Stern et al., 2019) to parallelly generate the remaining words.

3.1 Concept Graph Construction

Inspired by Xu et al. (2020a), we build a concept graph with two steps: vertex construction[2] and edge construction. Given a dialogue corpus S, we exploit a rule-based keyword extraction method to identify salient keywords from utterances in S (Tang et al., 2019). All extracted keywords are collected as dialogue concepts that represent vertices in the concept graph. For edge construction, we use pointwise mutual information (PMI) (Church and Hanks, 1989) to construct a concept pairwise matrix that characterizes the association between concepts in the observed dialogue data (Mou et al., 2016; Tang et al., 2019), where each concept pair consists of two concepts that are extracted from the context and the response, respectively. For each head vertex v^h, we select concepts with top PMI scores as tail vertices v^t and build edges by connecting v^h with all v^ts. In this way, we filter out low-frequency edges to narrow the search space for downstream concept planning.

3.2 Multi-Concept Planning Module

Given the dialogue context D, the historical concept flow F, and a concept graph $\mathcal{G}$, the goal of multi-concept planning is to predict a sequence of target concepts C, namely $P(C|D, F, \mathcal{G})$. All target concepts are extracted from $\mathcal{G}$ and arranged in a sequence $C = \{c_1, c_2, ..., c_t\}$, which reflects the order of target concepts in the final response.

Hierarchical Dialogue Encoder. To facilitate the understanding of dialogue context D, we employ Transformer blocks (Vaswani et al., 2017) to hierarchically encode dialogue context, aiming to capture the global semantic dependency between utterances. Formally, given the dialogue context $D = \{u_1, u_2, ..., u_N\}$ with N utterances, where

[2]The original constructed vertices in Xu et al. (2020a) involve what-vertices and how-vertices, where how-vertices represent different ways of expressing response content with a multi-mapping model (Chen et al., 2019). Here, we only collect what-vertices as dialogue concepts.

$u_i = \{w_{i1}, w_{i2}, ..., w_{in}\}$ is the word sequence of i-th utterance, we transform u_i into a sequence of hidden vectors with a Transformer encoder:

$$[\hat{\mathbf{h}}_i^{cls}, \hat{\mathbf{h}}_{i1}, ..., \hat{\mathbf{h}}_{in}] = \mathrm{TF}_{\theta_w}([\mathbf{e}_i^{cls}, \mathbf{e}_{i1}^{w}, ..., \mathbf{e}_{in}^{w}]). \tag{1}$$

Here, $\mathbf{e}_{ij}^{w}$ is the embedding of the j-th word in u_i. $\hat{\mathbf{h}}_i^{cls}$ and $\mathbf{e}_i^{cls}$ represent a special token [CLS] that is used to aggregate sequence representations, which is inspired by Devlin et al. (2019). Then, we collect utterance representations derived from [CLS] and input them into another Transformer encoder to hierarchically fuse context information:

$$[\mathbf{h}_1^{cls}, \mathbf{h}_2^{cls}, ..., \mathbf{h}_N^{cls}] = \mathrm{TF}_{\theta_u}([\hat{\mathbf{h}}_1^{cls}, \hat{\mathbf{h}}_2^{cls}, ..., \hat{\mathbf{h}}_N^{cls}]). \tag{2}$$

$\mathbf{h}_i^{cls}$ is a context-aware utterance representation that can be used to guide concept selection in the following steps.

Concept Flow Encoder. Formally, a concept flow $F = \{f_1, f_2, ..., f_N\}$ represents the observed concepts in the dialogue context, where f_i means a concept set corresponding to the i-th utterance that collects all the concept words in u_i, namely $f_i = \{c_{i1}, c_{i2}, ...c_{im}\}$. Here, c_{ij} is the j-th concept word in u_i. For an empty set $f_i = \emptyset$, a special NULL token is served as the concept word.

To capture information of history concept transitions, we exploit a vanilla GRU unit (Cho et al., 2014) to recursively read concept words in the flow:

$$\mathbf{s}_i = \mathrm{GRU}(\mathbf{s}_{i-1}, \mathbf{f}_i), \quad i \in [1, N]. \tag{3}$$

Here $\mathbf{f}_i$ denotes the representation of concept set f_i, which is calculated as a weighted sum of concept word embeddings $\mathbf{e}_{ij}^{c}$:

$$\begin{aligned} \mathbf{f}_i &= \sum\nolimits_{j=1}^{m} \alpha_{ij}^{f} \mathbf{e}_{ij}^{c}, \\ \alpha_{ij}^{f} &= \frac{\exp(\beta_{ij}^{f})}{\sum_{k=1}^{m} \exp(\beta_{ik}^{f})}, \\ \beta_{ij}^{f} &= \mathbf{s}_{i-1}^{\top} \mathbf{W}_f \, \mathbf{e}_{ij}^{c}, \end{aligned} \tag{4}$$

where $\mathbf{W}_f$ is trainable parameters. The output state $\mathbf{s}_{i-1}$ at the $i-1$ step is used as a query to compute β_{ij}^{f} scores, which can measure the preference of transitions to associated concepts. Empirically, $\mathbf{s}_0$ is a zero vector to initialize the recurrent process, and the final output $\mathbf{s}_N$ can serve as a memory to enable history-aware concept planning.

Multi-Concept Extractor. Recall that our goal is to produce a concept sequence C, which is a subsequence of the target response. Inspired by *pointer network* (Vinyals et al., 2015), we design a multi-concept extractor to achieve this goal, which can attentively read the dialogue context and the concept flow to sequentially extract target concepts from the contextually related subgraphs in $\mathcal{G}$.

To implement concept extraction in a sequential decoding manner, we use a Transformer decoder and compute its decoding states as follows:

$$\begin{aligned} \mathbf{H}^{cls} &= [\mathbf{h}_1^{cls}, \mathbf{h}_2^{cls}, ..., \mathbf{h}_N^{cls}], \\ \mathbf{m}_t &= \mathrm{TF}_{\theta_d}([\mathbf{e}_{1:t-1}^{c}], \mathbf{H}^{cls}). \end{aligned} \tag{5}$$

The utterance representations $\mathbf{H}^{cls}$ are memories for decoder-encoder attention. $[\mathbf{e}_{1:t-1}^{c}]$ denotes the embeddings of previously decoded concepts. $\mathbf{m}_t$ is the output state at step t conditioned on the dialogue context and partially decoded outputs.

Given the decoder state $\mathbf{m}_t$ and the concept flow memory $\mathbf{s}_N$, the following step is to select target concepts from $\mathcal{G}$. We first retrieve a group of subgraphs that corresponds to the concept set f_N of the last utterance u_N to prepare for the next round of concept transition. Here, each subgraph g_j consists of a hit concept $c_{Nj} \in f_N$ and its concept neighbours. Formally, $g_j = \{(c_j^{head}, c_{jk}^{tail})\}_{k=1}^{N_{g_j}}$, where c_j^{head} and c_{jk}^{tail} represent head concept vertex and tail concept vertex, respectively. N_{g_j} means the number of vertex pairs in g_j. Then, we employ a dynamic graph attention mechanism to calculate subgraph vectors $\mathbf{g}_j$ at each decoding step t to fuse information of all concept neighbours:

$$\begin{aligned} \mathbf{g}_j &= \sum\nolimits_{k=1}^{N_{g_j}} \alpha_{jk}^{g} [\mathbf{e}_j^{head}; \mathbf{e}_{jk}^{tail}], \\ \alpha_{jk}^{g} &= \frac{\exp(\beta_{jk}^{g})}{\sum_{l=1}^{N_{g_j}} \exp(\beta_{jl}^{g})}, \\ \beta_{jk}^{g} &= (\mathbf{W}_q^{g} [\mathbf{m}_t; \mathbf{s}_N; \mathbf{e}_j^{head}])^{\top} \cdot (\mathbf{W}_k^{g} \mathbf{e}_{jk}^{tail}). \end{aligned} \tag{6}$$

$\mathbf{W}_q^{g}$, $\mathbf{W}_k^{g}$ are trainable parameters. $\mathbf{e}_j^{head}$, $\mathbf{e}_{jk}^{tail}$ are embeddings of head and tail concepts in g_j. Here, α_{jk}^{g} is the probability of choosing c_{jk}^{tail} from all concept neighbours in g_j at step t conditioned on the dialogue context and the concept flow. We then compute the probability of choosing g_j at step t as a top-level concept selection, denoted as α_j^{t}:

$$\begin{aligned} \alpha_j^{t} &= \frac{\exp(\beta_j^{t})}{\sum_{l=1}^{m} \exp(\beta_l^{t})}, \\ \beta_j^{t} &= (\mathbf{W}_q^{t} [\mathbf{m}_t; \mathbf{s}_N])^{\top} \cdot (\mathbf{W}_k^{t} \mathbf{g}_j), \end{aligned} \tag{7}$$

where $\mathbf{W}_q^t$ and $\mathbf{W}_k^t$ are trainable parameters. Finally, the selection probability of target concepts at step t can be derived as:

$$P(c_t|c_{1:t-1}, D, F, \mathcal{G}) = \alpha_j^t \alpha_{jk}^g. \quad (8)$$

The multi-concept extractor has two stop conditions: 1) We add a special token c^{stop} to the concept neighbour set of g_j[3]. The extractor treats c^{stop} as a legal candidate target, and the selection of c^{stop} results in a stop action. 2) the number of target concepts exceeds N_{max}. Furthermore, for all the concepts extracted at step $k(k < t)$, we set their probabilities to 0 to avoid duplicate extraction.

3.3 Concept-Guided Insertion Transformer

After obtaining the target concept sequence C, the next step is to generate a response that covers C. General autoregressive approaches cannot ensure the success of introducing certain contents because they prefer to generate generic words (Mou et al., 2016). Given a substantially big language model, the problem might be alleviated but still cannot be completely solved. To address the issue, we use an Insertion Transformer (Stern et al., 2019) to generate a response based on C, which ensures the appearance of target concepts. On the other hand, the explicit planned concepts can be regarded as a prompt or a signal to guide the generation process. Generation is accomplished by repeatedly making insertions into a sequence initialized by C until a termination condition is met. At each decoding step t, the Insertion Transformer produces a joint distribution over the choice of words w_t and all available insertion locations $l_t \in [0, |\hat{y}_{t-1}|]$ in the previously decoded response $\hat{y}_{t-1}$:

$$\begin{aligned} &\hat{y}_0 = C, \\ &\hat{\mathbf{E}}_t = [\mathbf{e}_k^w | w_k \in \hat{y}_t], \\ &p(w_t, l_t | D, \hat{y}_{t-1}) = \mathrm{InsTF}(\mathbf{H}^{cls}, \hat{\mathbf{E}}_{t-1}), \end{aligned} \quad (9)$$

where $\hat{\mathbf{E}}_t$ is the word embedding list of $\hat{y}_t$. Notably, $\hat{y}_t$ has multiple available insertion locations, and we can perform parallel decoding by applying insertions at multiple locations simultaneously. For more details of Insertion Transformer, please refer to Stern et al. (2019) due to the space limitation.

[3]In this case, we make sure that $N_{g_j} > 0$, where g_j has at least one special vertex pair (c_j^{head}, c^{stop}).

	Persona	Weibo
# training pairs	101,935	1,818,862
# validating pairs	5,602	9,187
# testing pairs	5,317	9,186
# concept vertices in $\mathcal{G}$	2,409	4,000
# transition edges in $\mathcal{G}$	50,744	74,362
# concepts in each utterance	2.56	1.61

Table 1: Statistics of the dialogue datasets and the constructed concept graphs.

3.4 Training and Loss Functions

Given the list of ground truth concepts C in the target response y, the concept extractor is trained as a usual sequence generation model to minimize the negative log likelihood (NLL) loss as follows:

$$\mathcal{L}_C = \frac{1}{|C|} \sum_{t=1}^{|C|} -\log p(c_t|c_{1:t-1}, D, F, \mathcal{G}). \quad (10)$$

To train the Insertion Transformer, we first sample a subsequence $\hat{y}$ containing all the target concepts from the target response y. Then, for each of the $k+1$ locations $l = 0, 1, ..., k$ in $\hat{y}$, let $(w_{i_l}, w_{i_l+1}, ..., w_{j_l})$ be the span of words from the target response yet to be produced at location l. The loss function is finally defined as follows:

$$\mathcal{L}_R = \frac{1}{k+1} \sum_{l=0}^{k} \sum_{i=i_l}^{j_l} -\log p(w_i, l|D, \hat{y}) \cdot w_l(i). \quad (11)$$

Here $w_l(i)$ is a softmax weighting policy (Stern et al., 2019) that performs a weighted sum of the negative log-likelihoods of the words in the span. It encourages the generator to produce the central words of the span for a faster decoding process.

4 Experimental Settings

4.1 Datasets

Experiments are conducted on two public open-domain dialogue datasets **Persona-Chat** (Zhang et al., 2018) and **Weibo** (Shang et al., 2015). For Persona-Chat, the associated persona information is discarded so that the model can focus on the development of dialogues. Following previous works (Tang et al., 2019; Xu et al., 2020a), we employ a rule-based method to automatically extract concept words of each utterance, which combines tf-idf and POS features for scoring word salience. After dataset cleaning, we re-split the Persona-Chat dataset into train/valid/test sets as done in Tang et al. (2019), while the Weibo dataset is split in

Model	Persona-Chat					Weibo				
	BLEU	RG-1	RG-L	Dist-1	Dist-2	BLEU	RG-1	RG-L	Dist-1	Dist-2
Without Concept Planning										
Seq2seq+Att	.0325	.1698	.1691	.0127	.0402	.0106	.0790	.1004	.0160	.0520
Transformer	.0285	.1565	.1553	.0181	.0738	.0287	.1175	.1480	.0186	.0898
HRED	.0332	.1781	.1785	.0136	.0410	.0151	.0830	.1029	.0145	.0730
ReCoSa	.0306	.1576	.1606	.0192	.0820	.0232	.0858	.1069	.0171	.0841
With Concept Planning										
Seq2BF	.0197	.1234	.1201	.0595	**.3058**	.0157	.1506	.1434	**.0447**	.1952
CCM	.0389	.1902	.1922	.0463	.1537	.0249	.1883	.1845	.0257	.1750
ConceptFlow	.0401	.2216	.2154	.0487	.1939	.0282	.2133	.2071	.0348	.1948
CG-nAR (ours)	**.0477**	**.2611**	**.2502**	**.0626**	.2516	**.0304**	**.2576**	**.2417**	.0401	**.2809**

Table 2: Results of automatic evaluation for CG-nAR and baseline methods, which are categorized into two groups: *with / without* concept planning. The best results are highlighted in bold.

random. After constructing the graph of Persona-Chat, we randomly sample 100 concept vertices and 200 edges and ask three human annotators to evaluate their appropriateness. About 93% vertices and 72% edges are accepted by the annotators. For the graph of Weibo, we use the graph released by Xu et al. (2020a). Statistics of the two dialogue datasets along with the constructed graphs is shown in Table 1.

4.2 Comparison Methods

We compare CG-nAR with two groups of baselines: general seq2seq models and concept-guided systems. General seq2seq models produce responses conditioned on the dialogue messages without concept planning, including: **Seq2seq+Att** (Sutskever et al., 2014), a standard RNN model with attention mechanism; **Transformer** (Vaswani et al., 2017), a seq2seq model with a multi-head attention mechanism; **HRED** (Serban et al., 2016), a hierarchical encoder-decoder framework to model context utterances; **ReCoSa** (Zhang et al., 2019), a state-of-the-art model using the self-attention mechanism to measure the relevance of response and context. Concept-guided dialogue systems leverage concept information to control response generation, including: **Seq2BF** (Mou et al., 2016), a non-left-to-right generation model that explicitly incorporates a keyword into the response; **CCM** (Zhou et al., 2018a), a model that uses the graph attention mechanism to choose graph entities[4], and introduces them into response implicitly by a copy mechanism; **ConceptFlow** (Zhang et al., 2020), a state-of-the-art model that grounds each dialogue in the concept graph and traverses to distant concepts, which also generates concept words implicitly in an autoregressive manner; **CG-nAR** (our model), a model that explicitly introduces multiple concepts into responses with non-autoregressive generation.

[4]The original CCM uses an external knowledge graph. Here we adapt it to our constructed concept graph for a fair comparison. The same strategy is applied to ConceptFlow.

4.3 Implementation Details

We used VGAE (Kipf and Welling, 2016) to initialize the representation of concept vertices in the concept graph, and used Word2Vec (Mikolov et al., 2013) to initialize word embeddings. The embedding size of vertices and words was set to 128 and 300, respectively. We employed Adam (Kingma and Ba, 2015) with learning rate 1e-3 to train the concept extractor and the Insertion Transformer. All Transformer blocks have 3 layers, 768 hidden units, 8 heads, and the hidden size for all feed-forward layers is 2,048. The hidden size of GRU cells is 768. At inference time, the multi-concept extractor produces concepts greedily, and the maximum number of allowed concepts N_{max} was set to 5. For the Insertion Transformer, we used the configuration that achieved the best results reported in Stern et al. (2019). The whole model was trained for 100,000 steps with 8,000 warm-up steps on a 3090 GPU. Checkpoints were saved and evaluated on the validation set every 2,000 steps. Checkpoints with the top performance were finally evaluated on the test set to report final results.

5 Results and Analysis

5.1 Automatic Evaluation

We adopt widely used *BLEU* (Papineni et al., 2002) and *ROUGE* (Lin, 2004) to measure the relevance between the generation and the ground-truth. We report averaged BLEU scores with 4-grams at most and ROUGE-1/L (RG-1/L) F-scores. To measure the diversity of generated responses, we report

Model	App.	Inf.
Persona-Chat		
CG-nAR vs. ReCoSa	72.5%†	59.3%†
CG-nAR vs. Seq2BF	63.0%†	53.1%
CG-nAR vs. CCM	60.5%†	55.0%
CG-nAR vs. ConceptFlow	58.2%†	54.4%
Weibo		
CG-nAR vs. ReCoSa	78.2%†	61.9%†
CG-nAR vs. Seq2BF	65.5%†	52.9%
CG-nAR vs. CCM	61.4%†	54.0%
CG-nAR vs. ConceptFlow	61.1%†	54.7%

Table 3: Results of manual evaluation with *appropriateness* (App.) and *informativeness* (Inf.). The score is the percentage of times CG-nAR is chosen as the better in pairwise comparison with its competitor. Results marked with † are significant (using sign test, p<0.05).

Model	P	R	F1	Num.
Persona-Chat				
ReCoSa	.0137	.0077	.0099	1.29
Seq2BF	.0280	.0189	.0226	1.55
CCM	.2406	.1853	.2094	1.34
ConceptFlow	.3580	.4041	.3797	1.50
CG-nAR	**.5330**	**.5029**	**.5175**	2.17
Weibo				
ReCoSa	.0643	.0611	.0626	1.53
Seq2BF	.1685	.1657	.1671	1.58
CCM	.2514	.2059	.2264	1.61
ConceptFlow	.3859	.4177	.4012	1.78
CG-nAR	**.5119**	**.6455**	**.5710**	2.03

Table 4: Results of **Concept-P/R/F1** that compare the concepts in output responses with those in ground-truth ones. **Num.** denotes the average number of concepts predicted in output responses.

the ratio of distinct uni/bi-grams (*Dist-1/2*) in all generated responses (Li et al., 2016).

Table 2 shows the results of automatic evaluation for CG-nAR and baseline methods. All methods can be categorized into two groups: traditional seq2seq based generators and concept grounded methods. CG-nAR outperforms all other baselines significantly on BLEU and ROUGE scores (using Wilcoxon signed-rank test, p< 0.05), which manifests that the responses generated by CG-nAR match better with the ground-truth responses. This means CG-nAR can maintain the dialogue flow on-topic by the multi-concept planning mechanism. In terms of Dist-1/2 that measures the response diversity, all methods with concept planning can produce more diverse responses than those without, which indicates the problem of generic responses is alleviated by integrating concept information. Compared to the baselines with concept planning, CG-nAR has a better performance on response diversity. It verifies the effectiveness of our multi-concept planning module and the concept-guided non-autoregressive strategy, which can produce and combine multiple context-related concepts to compose diverse responses and keep concept words in the output response in an explicit manner.

5.2 Manual Evaluation

Considering automatic metrics may not suitably reflect the content to be evaluated, we further performed manual evaluation following previous works (Zhou et al., 2018a; Wu et al., 2020). Specifically, we randomly sampled 200 testing pairs from each test set and employed three annotators with professional background knowledge to evaluate the responses. Given a dialogue message, annotators were required to conduct pair-wise comparison between the response generated by CG-nAR and the one by a baseline (1,600 comparisons with four baselines on two datasets in total). For each comparison, annotators decided which response is better in terms of *appropriateness* (the model's ability to produce a fluent, coherent, and context-relevant response) and *informativeness* (if the response provides diverse information). For appropriateness, the percentage of pairs that at least 2 annotators gave the same judge (2/3 agreement) is 95.8%, and the percentage for 3/3 agreement is 62.7%. For informativeness, the at least 2/3 agreement is 89.0% and 3/3 agreement is 56.2%.

We compare CG-nAR against four baselines on Persona-Chat and Weibo (see Table 3). The score represents the percentage of times CG-nAR is chosen as the better in pair-wise comparisons. For appropriateness, CG-nAR significantly outperforms all other baselines on two datasets (using sign test, $p < 0.05$). It means that CG-nAR can generate more context-relevant and coherent responses accepted by annotators, which validates the effectiveness of our multi-concept planning module. In terms of informativeness, the percentages that CG-nAR wins ReCoSa are noticeably higher than those against other baselines. It indicates that systems with a concept planning mechanism can produce more informative responses by content introducing.

5.3 Analysis of Multi-Concept Planning

To validate if the multi-concept planning module has the ability to extract context-relevant concepts and form a coherent dialogue, we calculate the precision, recall, and F1 score of predicted concepts against golden ones in responses (Concept-

Model	BLEU	RG-1	RG-L	Dist-1	Dist-2	Concept-P	Concept-R	Concept-F1
CG-nAR	**.0477**	**.2611**	**.2502**	**.0626**	.2516	.5330	**.5029**	**.5175**
w/o. concept planning	.0217	.1504	.1568	.0350	.1884	.2711	.2507	.2605
w/o. concept flow encoder	.0344	.2283	.2232	.0583	.1938	**.5778**	.3696	.4508
w/o. hierarchical encoder	.0327	.2130	.2009	.0468	**.2650**	.3242	.4250	.3678
w. gated controller (AR)	.0405	.2399	.2320	.0423	.1893	.3926	.4177	.4048

Table 5: Ablation study using automatic metrics on the Persona-Chat dataset. The best results are highlighted.

Model	Param.	total time (sec)	words/sec
Transformer	127.3M	71.36	672.1
ReCoSa	143.8M	65.27	766.9
CG-nAR	172.1M	25.99	1131.1

Table 6: Inference speed on the Persona-Chat test set. **Param.** denotes the number of parameters.

P/R/F1). We also record the average number of predicted concepts to measure the model's ability to introduce multiple concepts. From Table 4 we can observe that CG-nAR achieves a higher recall and F1 score against all baselines by a large margin, especially for ReCoSa and Seq2BF. It probes that our concept planning module can successfully extract more concepts relevant to the dialogue. This is also reflected in the number of predicted concepts, where CG-nAR produces more concept words than those methods with autoregressive generators, e.g., CCM and ConceptFlow. It indicates that the concept-guided generator can effectively keep the concept information in output responses using a non-autoregressive generation mechanism.

5.4 Ablation Study

We perform ablation studies to validate the effectiveness of each part of CG-nAR. Table 5 shows the results. One of the variants is a vanilla Insertion Transformer where the concept planning module is removed. The model performance unsurprisingly degrades by a large margin, because the model might produce generic responses without concept planning. After removing the concept flow encoder, the information of historical concept transitions is missing, which also leads to a performance drop. We further replace the hierarchical dialogue encoder with a vanilla Transformer encoder, the performance drop shown in Table 5 indicates that it is necessary to capture the context dependency information when performing dialogue modeling. To probe the effectiveness of the concept-guided non-autoregressive strategy, we replace the Insertion Transformer with a universal Transformer framework equipped with a gated controller as done in Zhang et al. (2020), where the generation probabilities are calculated over the word vocabulary and the set of selected concept words. Table 5 shows that with the autoregressive decoding strategy, the performance drop is significant. A possible explanation is that the appearance of some key concepts cannot be guaranteed by such an implicit concept-oriented generator, especially when the generator encounters concepts that are not frequently seen in the training set.

5.5 Speed Comparison

Our concept-guided non-autoregressive generation model shows not only the superiority on response quality, but also gives a significant speed-up at test time over the methods equipped with autoregressive generators. The results of speed comparison is shown in Table 6. For a fair comparison, we choose the baselines with a Transformer encoder-decoder framework, since our customized Insertion Transformer uses the same model components. The main advantage of the insertion-based generator at inference time is that we can predict words at different insertion locations simultaneously. From Table 6 we can see that CG-nAR achieves substantially test-time speed-up compared to the two autoregressive generators (up to 2.7x in total time and 1.6x in word generation rate) even when CG-nAR has more parameters[5].

5.6 Case Study

To compare different models intuitively, we show two dialogue cases of the Persona-Chat dataset with output responses in Table 7. We observe that CG-nAR can successfully output context-associated concepts, e.g., *grow vegetable* that is related to *garden*, and *singer* that is related to *country music*. Compared to other baselines, CG-nAR produces a response that is more coherent and relevant to the dialogue context, and shows a more

[5]Here we test the autoregressive baselines with a beam size of 3 (used for their best scores). Without beam-search, they have significantly worse results, so we do not compare speed-ups with that version.

Context	A: What do you do for work? B: No work just the hits that's all I need. A: I see, I am a volunteer at my local animal shelter. B: Well this good work. I am a veteran and I have a garden.	A: Hi. How are you doing? B: I'm good, just finished practicing the guitar. You? B: Do you do that for a living? A: No, just a hobby because country music is my favorite.
Ground Truth	Thank you for your service! Do you grow any vegetables?	Who is your favorite singer?
ReCoSa	That's cool. Do you have any pets?	I'm a teacher. I've a dog.
Seq2BF	Cool! I work part time at an animal shelter.	I like music. I love my job.
CCM	I'm a teacher. What do you do?	Do you have any hobbies?
ConceptFlow	I enjoy to eat organic foods.	I love music. What are your hobbies?
CG-nAR	Do you grow vegetable for a living?	That's cool. Who is your favorite singer?

Table 7: Dialogue cases with output responses from different systems. Words in Blue are the observed concepts in the dialogue flow. Words in Red represent the context-associated concepts in the output response.

natural transition of concepts, which again proves the effectiveness of our concept-guided non-AR strategy for controllable dialogue generation.

6 Conclusion

In this work, we propose a novel concept-guided non-autoregressive approach for open-domain dialogue generation. It consists of a multi-concept planning module that selects multiple context-relevant concepts to facilitate a coherent dialogue, and a customized Insertion Transformer that produces a response based on the selected concepts to control the generation process. The experimental results show that our method can not only produce high-quality responses, but can also significantly speed up the inference time.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by China National Key R&D Program (No. 2017YFB1002104), National Natural Science Foundation of China (No. 61976056, 62076069), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103).

References

Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019. Generating multiple diverse responses with multi-mapping and posterior mapping selection. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4918–4924. ijcai.org.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

ChengYue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. FRAGE: frequency-agnostic word representation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1341–1352.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jiatao Gu, Qi Liu, and Kyunghyun Cho. 2019. Insertion-based decoding with automatically inferred generation order. *Transactions of the Association for Computational Linguistics*, 7:661–676.

Qinghong Han, Yuxian Meng, Fei Wu, and Jiwei Li. 2020. Non-autoregressive neural dialogue generation. *arXiv preprint arXiv:2002.04250*.

Xinyu Hua and Lu Wang. 2020. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online. Association for Computational Linguistics.

Jungo Kasai, James Cross, Marjan Ghazvininejad, and Jiatao Gu. 2020. Non-autoregressive machine translation with disentangled context transformer. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5144–5155. PMLR.

Yerbolat Khassanov, Zhiping Zeng, Van Tung Pham, Haihua Xu, and Eng Siong Chng. 2019. Enriching rare word representations in neural language models by embedding matrix augmentation. *arXiv preprint arXiv:1904.03799*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *NIPS Workshop on Bayesian Deep Learning*.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Brussels, Belgium. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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

Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. 2019. FlowSeq: Non-autoregressive conditional sequence generation with generative flow. 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 4282–4292, Hong Kong, China. Association for Computational Linguistics.

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

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Jinghui Qin, Zheng Ye, Jianheng Tang, and Xiaodan Liang. 2020. Dynamic knowledge routing network for target-guided open-domain conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8657–8664.

Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8697–8704.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. 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 1577–1586, Beijing, China. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831, Online. Association for Computational Linguistics.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. 2019. Insertion transformer: Flexible sequence generation via insertion operations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5976–5985. PMLR.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324, Online. Association for Computational Linguistics.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2692–2700.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3766–3772. ijcai.org.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3351–3357. AAAI Press.

Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020a. Conversational graph grounded policy learning for open-domain conversation generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1835–1845, Online. Association for Computational Linguistics.

Jun Xu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che. 2020b. Knowledge graph grounded goal planning for open-domain conversation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9338–9345.

Lili Yao, Ruijian Xu, Chao Li, Dongyan Zhao, and Rui Yan. 2018. Chat more if you like: Dynamic cue words planning to flow longer conversations. *arXiv preprint arXiv:1811.07631*.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. ReCoSa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730, Florence, Italy. Association for Computational Linguistics.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 2031–2043, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4623–4629. ijcai.org.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018b. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.