Continue or SHIFT: Learning Conversational Patterns for Dialogue Generation

Shaoxiong Feng*
Beijing Institute of Technology

XuanCheng Ren*
Peking University

Kan Li
Beijing Institute of Technology

Xu Sun
Peking University

ABSTRACT

In dialogues, it is often the case that the response could be either relevant or irrelevant to the given conversation context, depending on the speaker's intention of either topic continuity or topic shift. However, this aspect of dialogues is less explored in existing generative dialogue systems, because the widely-used encoder-decoder-based attention models are built upon the assumption that the target sequence is and *must be* relevant to the source sequence. In this work, we propose the loose coupling approach (LCA) to directly address the learning of such conversational patterns, which includes two parts: a high-level information selecting mechanism that enables the model to ignore the context or the previous part of the response, and a sampling-based response guide mechanism that mitigates the exposure bias problem in training. Human and automatic evaluation results demonstrate the effectiveness of our approach on capturing and describing both topic shift and continuity.

KEYWORDS

Dialogue Patterns, Loose Coupling, Selection Mechanism, Gumbel-Softmax

ACM Reference Format:
Shaoxiong Feng, XuanCheng Ren, Kan Li, and Xu Sun. 2020. Continue or SHIFT: Learning Conversational Patterns for Dialogue Generation. In *Proceedings of AIIS'20*. ACM, New York, NY, USA, 4 pages. https://doi.org/TBA

1 INTRODUCTION

Generative dialogue models have been getting increasing attention [14, 18], which are typically based on the encoder-decoder models [16] with attention mechanisms [1]. For facilitating the generation of high-quality responses, various neural methods are recently proposed, including hierarchical representation and generation [12, 13, 19], self-attention based models [22], novel objectives [6, 24], and external knowledge [7, 20, 26].

*Both authors contributed equally to this research.

AIIS'20, July 30, 2020, Virtual Event, China

ACM ISBN TBA…$TBA
https://doi.org/TBA

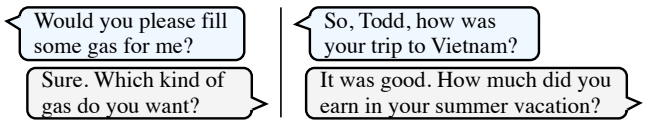

Figure 1: Examples of different patterns in dialogue. The response (left) continues the topic, while the response (right) shifts the topic.

However, unlike sequence transcription tasks such as machine translation, dialogue generation is a task where not only relevance matters but irrelevance also plays an important part. A variety of conversational patterns exist in dialogues, and there are intricate relations between response and context. Figure 1 shows two representative examples. If the response continues the topic in the context, it is usually relevant or dependent on a specific part of the context. If the response tries to shift the topic, it could be weakly relevant or even irrelevant to the context and the previous sentences in the response. The image captioning task also has the same phenomenon. Lu et al. [10] introduced a blank visual feature for attention when generating functional words that do not depend on images. Therefore, it is crucial to describe the degree of relevance and especially irrelevance of a response to its context for producing engaging dialogues. It is also important to notice that most of the existing sequence generation models [13, 17, 27] are built upon the assumption that the target sequence, i.e., the response, is and must be relevant to the source sequence, i.e., the context. The softmax function that is used to extract related words from sentences is inherently incapable of describing irrelevance. Hence, the information from the context and the previous response inevitably affects later responses, which may prevent the generation of responses with diverse conversational patterns.

To describe the irrelevance in dialogues and such kind of conversational patterns, we propose the loose coupling approach (LCA). Different from the conventional softmax function that is used to capture relevance, we propose to apply the sigmoid function to represent both relevance and irrelevance. On top of the current encoder and decoder, a high-level selecting mechanism is implemented to model the relations between a response and its context. Besides being conceptually simple, our experiments show that it is also very effective in describing topic continuity and topic shift. However, as the selecting mechanism is a kind of high-level information control, the model is subject to exposure bias [2] even more compared to low-level controls such as attention. To mitigate the side effect, we further propose the sampling-based response guide mechanism for enforcing the generation trajectory in inference close to the

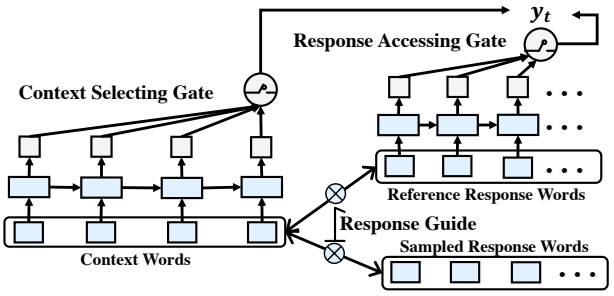

Figure 2: Illustration of the proposed approach.

reference in training on the semantic relation with the context. Concretely, the model is encouraged to minimize the difference between (a) the cosine similarity of the context and the reference and (b) that of the context and the sample generated by the model. Experimental results demonstrate that our model performs better than baselines on diversity and relevance.

2 APPROACH

In this section, we first describe the vanilla encoder-decoder framework, and then present the proposed loose coupling approach (LCA) shown in Figure 2.

2.1 Generative Dialogue Model

Generative dialogue systems generally consist of an encoder that extracts and represents information from the context and a decoder that generates the response based on the relevant context extracted by attention and the previous part of the response. Given a context of M words, $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$, the encoder maps it to source representation. The decoder predicts the next response word y_t using the context vector $\boldsymbol{c_t}$, which is obtained using attention to the source representation, and hidden state $\boldsymbol{h}_{t-1}$, which is the representation of the previous response words $\boldsymbol{y}_{1:t-1}$. Then the generation of each response word at timestep t could be denoted as:

$$\hat{\boldsymbol{v}}_t = \text{softmax}\left(\boldsymbol{W} f\left(\boldsymbol{h}_{t-1}, \boldsymbol{c}_t\right) + \boldsymbol{b}\right) \tag{1}$$

where $\hat{\boldsymbol{v}}_t$ is the logits, and f denotes a function that integrates the two kinds of information.

2.2 Context Selecting

The relation between the context and the response, especially in terms of relevance and irrelevance, is indicative of conversational patterns such as topic continuity and topic shift in dialogues. To explicitly model such relations, we introduce two gates to control the source of information.

The context does not always matter in response generation, especially when the response intends to change the topic. However, the attentive context vector $\boldsymbol{c}_t$ *always* contains information from context $\boldsymbol{x}$ based on $\boldsymbol{h}_{t-1}$, which means the decoder cannot ignore the context whether it is useful or not. It is because the attention mechanism based on the softmax function is only tasked to find the related context, without considering whether the context is actually needed. To mitigate this issue, we introduce the context selecting gate based on the sigmoid function:

$$\boldsymbol{g}_t = \sigma\left(\boldsymbol{W}_c \boldsymbol{c}_t + \boldsymbol{W}_h \boldsymbol{h}_{t-1}\right) \tag{2}$$

where σ is the sigmoid function limiting $\boldsymbol{g}_t$ to $[0, 1]$, and $\boldsymbol{W}_c, \boldsymbol{W}_h$ are parameters to be learned. The context selecting gate enables forgetting the entire conversational history.

It should be noted that since a response could contain multiple sentences, the relations among those sentences might also be different. For example, it is often observed that the response first answers the question in the context and then opens another topic to allow the conversation to proceed smoothly. Hence, we also introduce the response accessing gate that explicitly controls the information flow inside the response itself:

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{V}_c \boldsymbol{c}_t + \boldsymbol{V}_h \boldsymbol{h}_{t-1}\right) \tag{3}$$

where $\boldsymbol{V}_c$ and $\boldsymbol{V}_h$ are parameters to be learned.

Through the cooperation of the two gates, we can explicitly control the information flow from the available sources:

$$f(\boldsymbol{h}_{t-1}, \boldsymbol{c}_t) \triangleq \boldsymbol{a}_t = \boldsymbol{U}\left[\boldsymbol{g}_t \odot \boldsymbol{c}_t; \boldsymbol{f}_t \odot \boldsymbol{h}_{t-1}\right] \tag{4}$$

where $\odot$ means entrywise product, and $\boldsymbol{U}$ is the parameter to be learned that further integrates information from the sources. $\boldsymbol{a}_t$ serves as the final selected information for predicting next word. A byproduct of using σ as the gate function is that the values are continuous in the range $[0, 1]$, such that the model can trade off completely forgetting for completely remembering.

2.3 Response Guide

Since the selecting mechanism is a kind of high-level information control and the gate values affect more broadly than the attention weights, the model could be more likely to step into unseen trajectory, which exacerbates the exposure bias problem and prevents the learned patterns from taking effect. To tackle this, we propose to guide the response generation (as in inference) in the direction that resembles the reference response, and to constrain the inference from wandering too far from the familiar territory. By direction, we mean that the relation of the sampled response to the context should be similar to the relation of the reference to the context. Formally, the sampling-based reference guide mechanism is formulated as another learning objective:

$$L_G = \left(C\left(\boldsymbol{x}, \boldsymbol{y}\right) - C\left(\boldsymbol{x}, \hat{\boldsymbol{y}}\right)\right)^2 \tag{5}$$

where $\boldsymbol{x}, \boldsymbol{y}, \hat{\boldsymbol{y}}$ represents the context, the reference response, and the sampled response, respectively, and C is a semantic measure that evaluates to:

$$C(\boldsymbol{x}, \boldsymbol{y}) = \cos\left(e\left(\boldsymbol{x}\right), e\left(\boldsymbol{y}\right)\right) \tag{6}$$

where $e(\boldsymbol{x})$ is a vector representing the sequence $\boldsymbol{x}$ defined as $\sum_i w_i \boldsymbol{x}_i^{\text{emb}}$ based on the word embedding of either the encoder or the encoder, and the importance weight w_i that is the word frequency estimated from the training data. The proposed constraint is consistent with the motivation: if the reference pair $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is semantically far apart, then the pair with the sampled response $\langle \boldsymbol{x}, \hat{\boldsymbol{y}} \rangle$ should be correspondingly far apart, and vice versa.

However, simply implementing this process would result in a non-differentiable objective because of sampling operations. In order to achieve end-to-end gradient-based training, we adopt a continuous approximation of discrete sampling via Gumbel-softmax [4]. Specifically, let $\widetilde{\boldsymbol{v}}_t$ = Gumbel-softmax($\hat{\boldsymbol{v}}_t$, τ), where τ is the

DailyDialog	Dist-1 ↑	Dist-2 ↑	H^u_w ↑	H^b_w ↑	BLEU ↑	Average ↑	Greedy ↑	Extrema ↑	Relevance ↓	Continuity ↓	Shift ↓
LSTM+Att	0.0051	0.0167	8.0898	9.5449	0.2714	0.6276	0.4742	0.3130	2.818	2.409	2.500
Transformer	0.0114	0.0374	8.2345	10.0610	0.1672	0.6097	0.4570	0.2967	2.818	2.545	2.681
HRED	0.0051	0.0155	8.0878	9.8382	0.2064	0.6351	0.4873	0.3120	2.772	2.727	2.363
VHRED+BOW	0.0100	0.0316	7.7470	9.0858	0.1600	0.6316	0.4863	0.3115	2.909	3.090	2.909
LCA (LSTM+Att)	0.0160	0.0502	8.0106	9.7161	**0.2852**	**0.6473**	**0.4949**	**0.3150**	**2.227**	**2.181**	2.272
LCA (Transformer)	**0.0164**	**0.0600**	**8.2709**	**10.0716**	0.1674	0.6179	0.4546	0.2924	2.772	2.272	**2.181**
PersonaChat	Dist-1 ↑	Dist-2 ↑	H^u_w ↑	H^b_w ↑	BLEU ↑	Average ↑	Greedy ↑	Extrema ↑	Relevance ↓	Continuity ↓	Shift ↓
LSTM+Att	0.0005	0.0015	5.9044	5.7915	0.2820	0.6585	0.5563	0.2632	2.782	2.261	2.522
Transformer	0.0016	0.0046	6.5356	6.2018	0.2046	0.5634	0.4935	0.2352	3.043	2.652	2.521
HRED	0.0010	0.0023	6.5677	5.4994	0.1385	0.6015	0.4993	0.2716	2.565	2.608	2.130
VHRED+BOW	0.0010	0.0026	6.3954	5.6712	0.1649	0.6199	0.5156	0.2752	2.565	2.695	2.608
LCA (LSTM+Att)	0.0022	0.0076	6.3625	6.1104	**0.2874**	**0.6800**	**0.5621**	**0.2958**	**1.217**	**1.000**	**1.130**
LCA (Transformer)	**0.0030**	**0.0100**	**6.7999**	**6.5270**	0.2103	0.5842	0.5076	0.2406	2.391	2.434	2.130

Table 1: Results of automatic evaluation and human evaluation.

temperature. When $\tau \to 0$, $\widetilde{\boldsymbol{v}}_t$ becomes the one hot representation of token $\hat{\boldsymbol{y}}_t$. Then the word embedding $\hat{\boldsymbol{y}}^{\text{emb}}_t = \boldsymbol{E}\widetilde{\boldsymbol{v}}_t$, where $\boldsymbol{E}$ is the decoder word embedding matrix.

3 EXPERIMENT

We evaluate LCA on two dialogue datasets, i.e., DailyDialog [8], and Personachat [23], compared with re-implemented diverse baselines, i.e., **LSTM+Att** [16], upon which the proposal is implemented, **HRED** [15], which incorporates hierarchical representations of the context, **VHRED+BOW** [13, 25], which is based on variational auto-encoder with BOW loss, and **Transformer** [17], which only relies on the attention mechanism. Further introduction of the datasets, the baselines, and the training details, is provided in the appendix.

3.1 Experimental Results

Automatic Evaluation We adopt two kinds of automatic metrics. The reference-based metrics, **BLEU** [11] and embedding-based metrics [9], including embedding average (**Average**), embedding greedy (**Greedy**), embedding extrema (**Extrema**), are widely used to evaluate dialogue systems for *semantic relevance* [21, 25]. The count-based metrics, Distinct (**Dist-{1,2}**) [5] and Word Entropy [13], are used to evaluate the *lexical diversity* and the *information content* of the responses [3, 24]. We report the unigram and bigram version of Word Entropy, i.e., $\boldsymbol{H^u_w}$ and $\boldsymbol{H^b_w}$. As summarized in Table 1, we can see that LCA is capable of enhancing lexical diversity and information context without degrading semantic relevance. It is also interesting to see that in general models based on RNNs perform better on reference-based metrics, while Transformers are better at count-based metrics and produce responses that are more diverse and informative, which can be attributed to the self-attention that extracts relevant information more accurately.

Human Evaluation We conduct human evaluation in terms of topic relevance (**Relevance**), topic continuity (**Continuity**), and topic shift (**Shift**), which reflect whether the response is coherent with the context, whether the response further advances the current discussion after addressing the context, and whether the response introduces a new topic for the continuation of the conversation, respectively. Note that the new topic in **Shift** does not show any contradictory semantic with the context in comparison to the irrelevant topic in **Relevance**. For each dataset, 200 test examples are randomly selected, and three annotators are asked to rank the generated responses for each example. Ties are allowed. The individual ranks are then averaged to compose the final rank, and lower is better. The inter-annotator agreement in terms of Spearman's correlation coefficient for Relevance, Continuity, and Shift is 0.233, 0.252, and 0.294, which are all statistically significant ($p < 0.001$). As shown in Table 1, it is clear that LCA brings consistent improvements for the applied models, i.e., LSTM+Att and Transformer, which achieves the best result across the board, especially on PersonaChat.

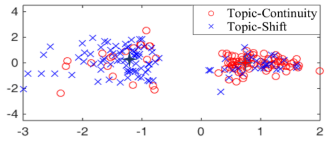

Figure 3: Visualization of pattern clustering. It verifies that the gate values learned by high-level information selecting mechanism are indicative of conversational patterns.

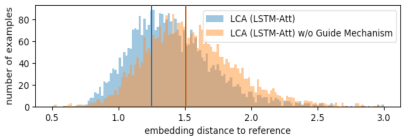

Figure 4: The distribution of embedding distance to reference. It demonstrates that with the guide mechanism the distribution of the generated responses is closer to the distribution of the reference responses.

3.2 Experimental Analysis

In this section, we conduct further analysis to prove the effectiveness of our method. Unless otherwise stated, the results are based on LSTM+Att and the validation set of DailyDialog.

Effect of Selecting Mechanism To further confirm the reliability and validity of our approach, we investigate the correlation between conversational patterns and learned gate values in a quantitative manner. The gate values for each example are first processed and then projected to 2 dimensions using principal component analysis (PCA). Each example is annotated as Topic-Continuity and Topic-Shift. Please refer to the appendix for the detailed settings. The results are shown in Figure 3. It is more than clear the examples

Input: I prefer well aged and fermented. I could bathe in it! // Good choice. I always like a nice dry white wine. // I think i should go grab a bottle now and get some dancing music on.
LSTM+Att: Do you have any pets? i do not have any pets. **Transformer:** I am not sure what you want to do? **HRED:** Do you have any hobbies? **VHRED:** What do you do for a living? **LCA (LSTM+Att):** That sounds nice. I like to listen to music while i am in the mood. **LCA (Transformer):** I like to listen to music. I love country music.
Input: No, we recently purchased a new house, so we can not afford it. Have you? // Yes i love mickey mouse such a cute little rat. // I enjoy going to concerts, i see the rolling stones every year.
LSTM+Att: I am a student. I am a student. **Transformer:** That is great. I am not sure. **HRED:** I am a huge fan of it. **VHRED:** That is cool. I like to go to the beach. **LCA (LSTM+Att):** Do you have any pets? I have a dog and i love to go out with my friends. **LCA (Transformer):** That is cool. I like to go to the park and i do not like to cook. Do you like sports?

Table 2: Examples of the generated responses.

form two distinct clusters, corresponding to topic continuity and topic shift, respectively. It proves that the selecting mechanism, albeit conceptually simple, is effective in capturing such kinds of conversational patterns.

Effect of Guide Mechanism In order to assess whether the guide mechanism mitigates the exposure bias problem, we introduce the embedding distance by calculating the euclidean distance between the sentence embedding of generated responses and references. From Figure 4, we can see that compared to only adopting the selecting mechanism, applying the guide mechanism can force the generated responses closer to the references, which in turn means the generated responses are more diverse and relevant. We also report all metrics results of the model without guide mechanism in the appendix, which show overall improvements, to further verify the above analysis.

Case Study We also provide some cases of the generated response in Table 2, in which the responses generated by our model show diverse conversational patterns. Specifically, LCA enables the model to generate responses that not only are specific to the context but also continues or shift the topic, as shown by the first and the second example, respectively. In contrast, the responses produced by baselines could be irrelevant to the context or hard to understand.

4 CONCLUSIONS

Different from machine translation, where the input and the output are tightly coupled in semantics, dialogue generation is a loose coupling task, where responses could be relevant and irrelevant to context, suggesting different patterns, i.e., topic continuity and topic shift. In this work, we propose the loose-coupling approach (LCA), which incorporates the selecting mechanism to control the high-level information flow from the context or the previous part of the response to allow the modeling of irrelevance. Furthermore, a guide mechanism is implemented to direct the response sampling in inference to mitigate the exposure bias problem. Experimental results validate the effectiveness of our approach in terms of describing and improving both topic shift and topic continuity.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

[2] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *NIPS*. 1171–1179.

[3] Richard Csaky, Patrik Purgai, and Gábor Recski. 2019. Improving Neural Conversational Models with Entropy-Based Data Filtering. In *ACL (1)*. 5650–5669.

[4] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *ICLR (Poster)*.

[5] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *HLT-NAACL*. 110–119.

[6] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *EMNLP*. 1192–1202.

[7] Xiang Li, Lili Mou, Rui Yan, and Ming Zhang. 2016. StalemateBreaker: A Proactive Content-Introducing Approach to Automatic Human-Computer Conversation. In *IJCAI*. IJCAI/AAAI Press, 2845–2851.

[8] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP(1)*. 986–995.

[9] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*. 2122–2132.

[10] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In *CVPR*. 3242–3250.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*. 311–318.

[12] 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 *AAAI*. 3776–3784.

[13] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *AAAI*. 3295–3301.

[14] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *ACL (1)*. 1577–1586.

[15] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *HLT-NAACL*. 196–205.

[16] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS*. 3104–3112.

[17] 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 *NIPS*. 5998–6008.

[18] Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. *CoRR* abs/1506.05869 (2015).

[19] Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. 2018. Hierarchical Recurrent Attention Network for Response Generation. In *AAAI*. 5610–5617.

[20] Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2017. Incorporating loose-structured knowledge into conversation modeling via recall-gate LSTM. In *IJCNN*. 3506–3513.

[21] Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018. Reinforcing Coherence for Sequence to Sequence Model in Dialogue Generation. In *IJCAI*. 4567–4573.

[22] 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 *ACL (1)*. 3721–3730.

[23] 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 *ACL (1)*. 2204–2213.

[24] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization. In *NeurIPS*. 1815–1825.

[25] Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *ACL (1)*. 654–664.

[26] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI*. 4623–4629.

[27] Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective Encoding for Abstractive Sentence Summarization. In *ACL (1)*. Association for Computational Linguistics, 1095–1104.