Topic Enhanced Controllable CVAE for Dialogue Generation (Student Abstract)

Yiru Wang,[1] **Pengda Si,**[1] **Zeyang Lei,**[*2] **Yujiu Yang** [†1]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, P. R. China
[2]Baidu Inc., Beijing, P. R. China
{wangyiru17, spd18}@mails.tsinghua.edu.cn, zeyanglei@gmail.com, yang.yujiu@sz.tsinghua.edu.cn

Abstract

Neural generation models have shown great potential in conversation generation recently. However, these methods tend to generate uninformative or irrelevant responses. In this paper, we present a novel topic-enhanced controllable CVAE (TEC-CVAE) model to address this issue. On the one hand, the model learns the context-interactive topic knowledge through a novel multi-hop hybrid attention in the encoder. On the other hand, we design a topic-aware controllable decoder to constrain the expression of the stochastic latent variable in the CVAE to reduce irrelevant responses. Experimental results on two public datasets show that the two mechanisms synchronize to improve both relevance and diversity, and the proposed model outperforms other competitive methods.

Introduction

Neural dialogue systems tend to generate monotonous and meaningless responses due to the information loss in the encoding phase, and the information provided only by the encoding vector obtained from the encoder for the decoding phase is far from sufficient. Most existing works introduce additional information such as context (Serban et al. 2016), keyword (Xing et al. 2017) or knowledge-base (Young et al. 2018) to improve response generation quality. Conditional variational auto-encoders (CVAEs) (Zhao, Zhao, and Eskénazi 2017) can generate informative responses relying on the variability of the stochastic latent variable. Although effective, many methods establish a trade-off between relevance and diversity. For example, in CVAE, the randomness of the latent variable can result in responses deviated from the conversation context. Traditional keyword methods are limited by the vocabulary size and the fixed keyword number, also not able to dynamically interact with the context information to learn the comprehensive topic knowledge.

To address the aforementioned limitation, we propose a novel topic enhanced controllable CVAE (TEC-CVAE) model, which aims to leverage the context-interactive topic knowledge and topic-aware controllable latent variable to further improve the relevance and informativeness of generated responses. More specifically, (1) in the encoder, we introduce a coarse keyword sequence with additional multi-hop hybrid attention layers. The interrelationship among keywords is learnt to fuse the topic information throughout the dialogue for each keyword representation. At the same time, a specific contextual topic vector is produced based on the interaction between the utterance and the keyword vectors above for final representation of each utterance. (2) In the decoder, we utilize an overall topic vector to generate the latent variable. Furthermore, we design a novel expression gate to dynamically control the influence of the randomness brought by the latent variable to avoid generating irrelevant responses. The experimental results show that our model is efficient and achieves state-of-the-art results.

Our Model

We extract nouns and named entities (NE) in the input dialogue histories as topic keywords. If no nouns and NE, we use adjectives and verbs instead. We then filter universal words according to mutual information, and finally get the keyword sequence $K = [k_1, ..., k_m]$. The encoder is a hierarchical network with sentence-level and context-level layers. The sentence-level encoder produces an encoding vector for each utterance, and obtains the sequence of utterance representations $U = [u_1, ..., u_n]$ for input dialogue histories.

Multi-hop Hybrid Attention

We first apply a self-attention layer on the keyword sequence K to obtain the representation r_i for k_i. Then, we calculate the attention of the utterance vector u_i to the keyword vectors to produce a contextual fused topic vector t_i for each utterance.

$$
\begin{aligned}
r_i &= \sum_{j=1}^{m} a_{ij}^{k} e_j, \text{ with } a_{ij}^{k} = \frac{\exp(\sigma([e_i, e_j])}{\sum_{p=1}^{m} \exp(\sigma([e_i, e_p])]} \\
t_i &= \sum_{j=1}^{m} a_{ij}^{u} r_j, \text{ with } a_{ij}^{u} = \frac{\exp(\sigma([u_i, r_j])}{\sum_{p=1}^{m} \exp(\sigma([u_i, r_p])]}
\end{aligned}
\tag{1}
$$

[*] This work was done when Zeyang Lei was in Tsinghua Univ.

[†] Corresponding author. This work was supported partially by Guangdong Basic and Applied Basic Research Foundation (No.2019A1515011387), and Shenzhen Basic Research Foundation (No.JCYJ201704is170118573).

where e_i is the word embedding of the keyword k_i. σ is a score function that defines the relation between two vectors as $\sigma([u_i, r_j]) = u_i \cdot (w r_j)^T / \sqrt{d_u}$, where d_u is the dimension of u_i, and w is the parameter vector to be learnt.

In this way, each keyword representation actually fuses the global coarse dialogue information. Furthermore, we concatenate u_i and t_i as the final representation $u_i^h = [u_i; t_i]$ of the i-th utterance, which compresses the context information into u_i^h via t_i. The context-level encoder receives the sequence $[u_1^h, ..., u_n^h]$ to generate the context vector c.

Topic-aware Controllable Decoder

We introduce the overall topic vector t^o by applying a mean-pooling on the keyword representations $t^o = \sum_{i=1}^{m} r_i/m$ into the latent variable z generation in CVAE. Specifically, the recognition network $q_\phi(z|y, c, t^o) \sim \mathcal{N}(\mu, \sigma^2 I)$ is with $[\mu; log(\sigma^2)] \triangleq W_r[y; c; t^o] + b_r$, where y is the encoding vector of the target response. The prior network $p_\theta(z|c, t^o) \sim \mathcal{N}(\mu', \sigma'^2 I)$ is with $[\mu'; log(\sigma'^2)] \triangleq MLP([c; t^o])$. The decoder is a GRU network with initial state $s_0 = W_i[z; c; t^o] + b_i$. For word prediction, we design an expression gate g^e in the GRU cell to dynamically control the strength of z. The decoding at time step t is as follows:

$$\begin{aligned}
&g_t^e = \sigma(W_x^e[y_t; t^o] + W_h^e s_{t-1} + W_z^e z) \\
&\tilde{s}_t = f(W_x[y_t; t^o] + W_h(g_t^r \circ s_{t-1}) + W_z(g_t^e \circ z)) \qquad (2) \\
&s_t = (1 - g_t^z) \circ s_{t-1} + g_t^z \circ \tilde{s}_t
\end{aligned}$$

where g^r and g^z are the reset gate and update gate of GRU. y_t is the t-th input word. σ is the $sigmoid$ function and f is the $tanh$ function. We allow the dynamically controllable z to participate in the generation at each decoding step.

Experiments

Datasets

All the experiments in this paper are carried out on two public datasets, one is the **DailyDialog** dataset collected by Li et al. (2017) and the other is the **OpenSubtitles** dataset from the OpenSubtitles website [1].

Baseline Methods

We compare our model with several state-of-the-art baseline methods, including the attention-based sequence-to-sequence model(S2SA) (Bahdanau, Cho, and Bengio 2015), HRED (Serban et al. 2016) and CVAE (Zhao, Zhao, and Eskénazi 2017).

Implementation Details

In all experiments, all the initial weights are sampled from a uniform distribution [-0.08, 0.08] and all the bias vectors are initialized to zero. The word embedding size is 200 and the topic vector size is 30. The sentence encoder, context encoder and response decoder are GRU with hidden size of 300, 600 and 400 respectively. The MLP in the prior network has a hidden size of 400, and the latent variable z has a size of 200. We use the Adam optimizer with batch size 30. The learning rate is 0.001 and the gradient clip is at 5.

	Dailydialog Corpus			
Method	**BLEU**	**Dist-1**	**Dist-2**	**KL cost**
S2SA	10.30	884/0.0134	2382/0.0495	-
HRED	11.36	872/0.0123	2399/0.0470	-
CVAE	14.99	2766/0.0269	14791/0.1876	17.32
MHA-CVAE	15.38	3005/**0.0295**	15879/0.2033	17.44
TACD-CVAE	15.63	3002/0.0275	16963/0.2035	18.02
TEC-CVAE	**16.15**	**3228**/0.0279	**18111/0.2040**	**18.41**
	OpenSubtitles Corpus			
S2SA	7.53	271/0.0031	648/0.0136	-
HRED	10.46	325/0.0033	704/0.0122	-
CVAE	12.16	3063/0.0228	20972/0.2242	14.19
MHA-CVAE	12.27	3461/0.0256	21446/0.2266	14.70
TACD-CVAE	**12.59**	4172/0.0297	24901/0.2491	15.04
TEC-CVAE	12.41	**4365/0.0315**	**25728/0.2618**	**16.09**

Table 1: Evaluation results. MHA-CVAE and TACD-CVAE represent that only multi-hop hybird attention or topic-aware controllable decoder is used in our model, respectively.

Experimental Results

We adopt BLEU and Dist-n to evaluate the relevance and diversity respectively. Dist-n are the number and proportion of distinct n-grams in all generated tokens. The experimental results are shown in Table 1. Compared to the baseline methods, our models achieve better results on both datasets. Higher KL cost indicates our model learns a more meaningful latent variable which benefits from the integrated topic knowledge. Two variant models are superior to baselines and TEC-CVAE gains the best overall performance. This verifies that both mechanisms contribute to the improvement of diversity and relevance simultaneously in response generation.

Conclusion

We propose a novel topic enhanced controllable CVAE model which learns the context-interactive topic knowledge and topic-aware controllable latent variable. Experimental results show that our method achieves better results in both relevance and informativeness of generated responses.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. *ICLR* abs/1409.0473.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, 986–995.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, 3776–3784.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W. 2017. Topic aware neural response generation. In *AAAI*, 3351–3357.

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

Zhao, T.; Zhao, R.; and Eskénazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, 654–664.

[1] https://www.opensubtitles.org