Dual Contrastive Learning for Unsupervised Knowledge Selection

Xiang Cao[1], Lei Wang[2] and Yujiu Yang[1]

[1]Tsinghua Shenzhen International Graduate School, Shenzhen, China
[2]Ping An Technology (Shenzhen) Co., Ltd., Shenzhen, China
cao-x19@mails.tsinghua.edu.cn, wangleis25@pingan.com.cn,
yang.yujiu@sz.tsinghua.edu.cn

Abstract

Although dialogue systems based on the Seq2Seq model have achieved success, they suffer from tending to generate general responses. Recent works have shown that selecting external knowledge is helpful for dialogue systems to generate informative and diverse responses. However, selecting appropriate knowledge from an unlabeled knowledge set, which is referred to as unsupervised knowledge selection, remains a tricky challenge. Therefore, we propose a dual contrastive method, which utilizes two source-target pairs which are based on the same knowledge set to construct dual contrasts. Specifically, for a source utterance, we consider its paired and unpaired target response as a positive and negative sample, then obtain the positive and negative posterior distribution over the knowledge candidates set, respectively. Then we lead the prior distribution to be close to the positive posterior distribution and distant from the negative one. Similarly, the posterior distribution is treated with the same criterion. Experimental results show that our method improves generated responses in terms of BLUE, DISTINCT, and knowledge utilization. Our codes are available at https://github.com/CaoXiang1997/DualCL4UKS.

1. Introduction

A dialogue system aims to produce an appropriate response given a post as its input. Recently, the generative method based on the encoder-decoder frame [4] has attracted considerable attention for generating fluent responses [15, 17]. However, it tends to generate general and boring responses like ”I don't know”, lending conversations into unattractive and boring situations.

Existing works have shown that leveraging external knowledge is helpful for dialogue systems to improve informativeness and diversity of generated responses [6, 10, 14]. External knowledge is sometimes in the form of multiple utterances in a set of knowledge [19, 5]. Intuitively, external knowledge builds an information bridge for dialogue systems from the source to the target. However, not all utterances in the knowledge set help models generate appropriate responses, and manual annotations are often expensive. Therefore, it is necessary to select appropriate knowledge from an unlabeled knowledge set, referred to as unsupervised knowledge selection.

For unsupervised knowledge selection, a tricky challenge is the lack of supervisory signals. Several works proposed attentive methods which use attention mechanism [10] or its more powerful varieties [14] to calculate the probability distribution over the knowledge set and select knowledge softly. However, those attentive methods only relied on the distant supervisory signal from the cross-entropy loss between the generated response and the target response to supervise knowledge selection, helpless to the lack of the supervisory signal. [10] attempted to use target response to compute posterior distribution as guidance for prior distribution, but still didn't provide a strong enough supervisory signal.

In this paper, inspired by previous works on contrastive learning [2, 7], we propose a dual contrastive method for unsupervised knowledge selection. From contrastive learning, the model benefits from the contrast between positive samples and negative samples. We think that appropriately selected knowledge is helpful for the model to distinguish positive samples from negative ones. Specifically, for a

DOI reference number: 10.18293/SEKE2022-054

source utterance, we consider its paired target response as the positive response and an unpaired target response as the negative response, compute positive and negative posterior distribution on the positive and negative response, respectively, and lead the prior distribution to be close to the positive posterior distribution and distant from the negative one. Similarly, we utilize the same criterion for the posterior distribution. Our contributions are listed as follows.

- As far as we know, we are the first to introduce contrastive learning into unsupervised knowledge selection, which provides an alternative solution for the label unavailability issue.
- We propose a dual contrastive method for unsupervised knowledge selection, which learns prior/posterior distribution from contrasts between positive and negative posterior/prior distributions.
- Experimental results show that our method outperforms other existing competitive methods on diversity and knowledge incorporation of generated responses and almost flats on other metrics.

2. Related Work

Sequence to sequence (Seq2Seq) models [4] promote the development of dialogue systems, but it suffers from generating general responses. Recently, some works utilize external knowledge to help dialogue systems generate diverse and informative responses. In some early works [6, 12], the model encoded external text entirely into a vector, which led to irrelevant knowledge noise in generated responses. Therefore, unsupervised knowledge selection became a research hotspot whereas the lack of labels. [10] proposed a prior-posterior framework that computes posterior distribution by the ground-truth response and drives the prior distribution to approach the posterior distribution. [14] proposed a global-to-local knowledge selection mechanism where the global knowledge selection module forms a topic transition vector and the local knowledge selection module select knowledge at each decoding step under the guidance of the topic transition vector. [16] proposed a teacher-student framework where the teacher builds response-aware document memory given the ground-truth response and the student learns response-anticipated document memory from the teacher. [3] introduced knowledge distillation to address the exposure bias issue of knowledge selection. [11] proposed recurrent knowledge interaction among decoding steps and introduced a knowledge copy mechanism to copy words from external knowledge. [8] proposed a sequential latent model which uses sequential latent to model the knowledge selection process in multi-turn dialogue generation.

However, existing works use one single source-target pair to select knowledge, ignoring the difference between knowledge selected by two source-target pairs. That results in more attention to the common information than the special information of knowledge candidates, which is bad for the model to incorporate relevant knowledge and generate diverse responses. In this paper, We propose a dual contrastive method for unsupervised knowledge selection, which leads the model to learn the difference in knowledge selection between two source-target pairs and generate responses based on specially selected knowledge.

3. Model

As presented in Figure 1, the architecture overview of our model is generally based on a sequence-to-sequence frame. We let $\left\{(x^i, y^i)\right\}_{i=1}^{n}$ denote a set of multi-turn dialogue, where n is the turn number, and x^i and y^i are the source utterance and the target utterance of the i-th turn, respectively. $K = \left\{k^j\right\}_{j=1}^{N}$ denotes a set of knowledge utterances, where N is the number of utterances in K, and k^j is the j-th knowledge utterance. Our model's goal is to select correct a knowledge utterance k^j from K and generate appropriate responses y^i for each x^i.

3.1. Base Architecture

Our method is based on an encoder-decoder architecture with a prior-posterior knowledge selector which uses the posterior distribution to guide the prior distribution.

Encoder We implement a source encoder with a bidirectional gated recurrent unit (GRU), which encode source utterance x^i into a forward hidden state $\overrightarrow{h}_x^i$ and a backward hidden state $\overleftarrow{h}_x^i$ for each x^i. We concatenate the last hidden states in two directions into h_x^i as the representation vector of x^i as follows.

$$\overrightarrow{h}_t = GRU(x_t^i, \overrightarrow{h}_{t-1}) \tag{1}$$

$$\overleftarrow{h}_t = GRU(x_t^i, \overleftarrow{h}_{t+1}) \tag{2}$$

$$h_x^i = [\overrightarrow{h}_{|x^i|}^i; \overleftarrow{h}_1^i] \tag{3}$$

where $[;]$ represents vector concatenation, $|x^i|$ represents the token number of x^i. In this way, we encode x^i as h_x^i for each i.

We implement a knowledge encoder with the same structure as the source encoder, but they don't share any parameters. Similarly, we concatenate the last hidden states of two directions into an overall vector h_k^j for each k^j. Moreover, we use the knowledge encoder to encode y^i as h_y^i for each i.

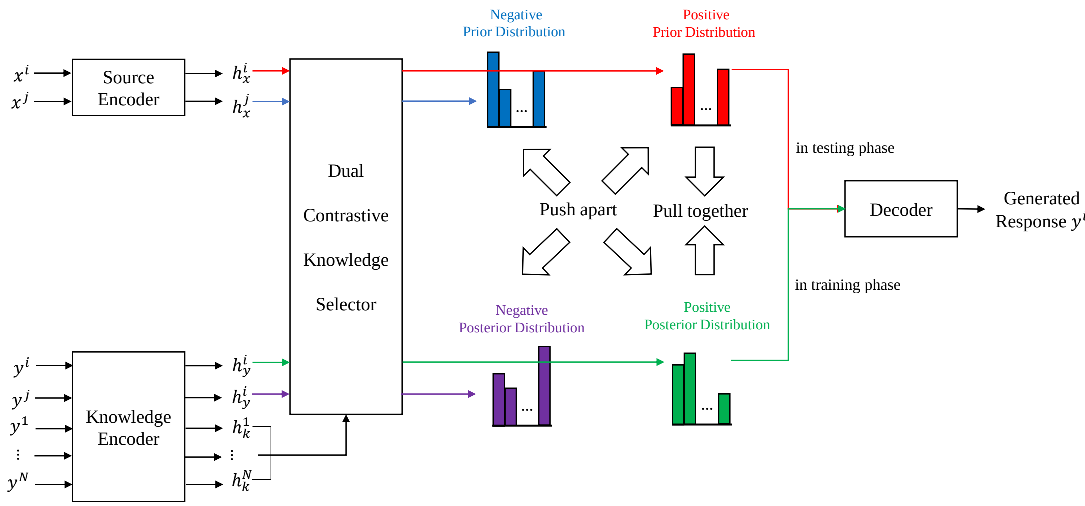

Figure 1. The architecture overview of our model. In this figure, we only show contrasts where the i-th turn (x^i, y^i) is the positive sample and the j-th turn (x^j, y^j) is the negative one.

Knowledge Selector The goal of the knowledge selector is to select appropriate knowledge from the knowledge set. Inspired by [10], we set two modes for the knowledge selector – the prior and posterior modes. In the prior mode, the knowledge selector computes a prior distribution p_x^i using the source utterance x^i. In the posterior mode, the knowledge selector computes a posterior distribution p_x^i using the target utterance y^i. The knowledge selector is implemented with attention mechanism [1] as follows.

$$p(k|z) = softmax(h_z \cdot [h_k^1, \cdots, h_k^n]) \tag{4}$$

where $z \in \{x^i\}_{i=1}^n \cup \{y^i\}_{i=1}^n$ is the representation vector of an source or target utterance. For convenience, we use p_x^i and p_y^i as simplifications of $p(k|x^i)$ and $p(k|y^i)$, respectively.

In the training phase, we use the posterior distribution as knowledge selection distribution to sample a knowledge utterance. In the testing phase, we have no choice but to use the prior distribution, because the target utterance is to be generated and can not be used to compute the posterior distribution.

To ensure the knowledge selector work in the testing phase, we use the posterior distribution to guide the prior distribution in the training phase. Therefore, we introduce the Kullback-Leibler Divergence (KLD) loss to minimize the distance between the prior distribution and the posterior distribution.

$$\ell_{KLD}^i = KLD(p_y^i || p_x^i) = p_y^i \log \frac{p_y^i}{p_x^i} \tag{5}$$

When the knowledge selection distribution $p(k)$is given, the selected knowledge utterance $sk \sim p(k)$ is sampled according to it.

Decoder The decoder integrates the selected knowledge h_{sk}^i and generates response word by word. We use a hierarchical gated fusion unit (HGFU) [18] to implement it. An HGFU consists of two GRUs, which are fed by the word generated in the last step y_{t-1} and the selected knowledge h_K, respectively, as follows.

$$\begin{aligned} s_{y,t}^i &= GRU(emb(y_{y,t-1}^i), s_{t-1}^i, c_t^i) \\ s_{k,t}^i &= GRU(h_{sk}^i, s_{t-1}^i, c_t^i) \end{aligned} \tag{6}$$

where emb is the embedding layer, s_{t-1}^i is the last hidden state of the decoder, c_t^i is the attentive context vector.

Then, HGFU combines the s_t^y and s_t^k with a soft gate g as follows.

$$s_t^i = g\dot{s}_{y,t}^i + (1-g) \odot s_{k,t}^i \tag{7}$$

g is computed by s_t^y and s_t^k through multilayer perceptrons and control the their contributions to the final hidden state s_t.

The word is sampled from a distribution computed by s_t^i and c_t^i as follows.

$$y_t^i \sim p_t^i = softmax(W_o[s_t^i; c_t^i]) \tag{8}$$

where W_o is the parameters of the output layer.

Loss Function We introduce the negative-log likelihood(NLL) loss to measure the difference between the response generated by the model and the target response as follows.

$$\ell^i_{NLL} = -\frac{1}{|y^i|}\sum_{t=1}^{|y^i|}\log p(y^i_t|y^i_{t-1}, x^i, sk^i) \tag{9}$$

Like [10], we introduce the bag-of-words(BOW) loss to ensure the accuracy of the selected knowledge as follows.

$$\ell^i_{BOW} = -\frac{1}{m}\sum_{t=1}^{m}\log p(y_t|k) \tag{10}$$

Therefore, the total loss function of our model is as follows.

$$\ell = \ell_{KS} + \ell_{NLL} + \ell_{BOW} \tag{11}$$

3.2. Dual Contrastive Knowledge Selector

We think that an appropriately selected knowledge is helpful not only for approaching the target response but also for distinguishing the target response from other responses. Therefore, we propose two kinds of contrastive loss as follows. For the posterior distribution p^i_y, we use the prior distribution p^i_x of the same turn as the positive and the prior distributions $\{p^j_x\}^n_{j=1,j\neq i}$ of other turns as the negatives. Then, we minimize the distance of p^i_y from the positive prior distribution p^i_x and maximize the average distance of p^i_y from the negative prior distributions $\{p^j_x\}^n_{j=1,j\neq i}$. In this way, We introduce the prior contrast and propose the prior contrastive loss as follows.

$$\begin{aligned}\ell^i_{PRIOR} &= p^i_y \log\frac{p^i_y}{p(k|x^i} - \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} p^i_y \log\frac{p^i_y}{p^j_x} \\ &= -p^i_y(\log p^i_x - \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\log p^j_x)\end{aligned} \tag{12}$$

To ensure ℓ_{PRIOR} is positive, we change it to its final form as follows.

$$\ell^i_{PRIOR} = -p^i_y[\log p^i_x + \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\log(1-p^j_x)] \tag{13}$$

Similarly, for the prior distribution p^i_x, we use the posterior distribution p^i_y of the same turn as the positive and the posterior distributions $\{p^j_y\}^n_{j=1,j\neq i}$ of other turns as the negatives. Then, we minimize the distance of p^i_x from the positive posterior distribution p^i_y and maximize the average distance of p^i_y from the negative posterior distributions $\{p^j_y\}^n_{j=1,j\neq i}$. In this way, we introduce the posterior contrast and propose the posterior contrastive loss as follows.

$$\begin{aligned}\ell^i_{POST} &= p^i_y \log\frac{p^i_y}{p(k|x^i} - \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} p^j_y \log\frac{p^j_y}{p^i_x} \\ &= [p^i_y\log p^i_y - \frac{1}{n-1}\sum_{j=1,j\neq i} p^j_y\log p^j_y] \\ &\quad - [p^i_y\log p^i_x - \frac{1}{n-1}\sum_{j=1,j\neq i} p^j_y\log p^i_x]\end{aligned} \tag{14}$$

For the whole multi-turn dialogue, we have

$$\sum_{i=1}^{n}[p^i_y\log p^i_y - \frac{1}{n-1}\sum_{j=1,j\neq i} p^j_y\log p^j_y] = 0 \tag{15}$$

Therefore, ℓ^i_{POST} can be simplified as follows.

$$\ell^i_{POST} = -p^i_y\log p^i_x + \frac{1}{n-1}\sum_{j=1,j\neq i} p^j_y\log p^i_x \tag{16}$$

Similarly, we change ℓ^i_{POST} to ensure it is positive as follows.

$$\ell^i_{POST} = -p^i_y\log p^i_x - \frac{1}{n-1}\sum_{j=1,j\neq i} p^j_y\log(1-p^i_x) \tag{17}$$

Overall, the total loss function of knowledge selection is as follows.

$$\ell^i_{KS} = \ell^i_{KL} + \alpha\cdot\ell^i_{PRIOR} + \beta\cdot\ell^i_{POST} \tag{18}$$

where α and β are coefficients to control the contribution of ℓ^i_{PRIOR} and ℓ^i_{POST}, respectively.

4. Experiments

4.1. Experiment Settings

Datasets We carry experiments on two open-domain knowledge-grounded dialogue datasets, namely PersonaChat [19] and Wizard-of-Wikipedia [5]. Although Wizard-of-Wikipedia has labels for knowledge selection, we did not use them because we focus on improvements in unsupervised knowledge selection.

Baselines We compared our models with the following baselines.

Table 1. Automatic Evaluation on PersonaChat and Wizard-of-Wikipedia.

Datasets	Models	BLEU-1 / 2 / 3	Distinct-1 / 2 / 3	Knowledge-R / P / F1
PersonaChat	Seq2Seq	0.1764 / 0.0725 / 0.0315	0.0136 / **0.1015** / **0.2908**	0.0062 / 0.0206 / 0.0095
	PostKS	0.1736 / 0.0720 / 0.0326	0.0136 / 0.0968 / 0.2676	0.0098 / 0.0407 / 0.0158
	our model	**0.1799** / **0.0739** / **0.0333**	**0.0144** / 0.1011 / 0.2804	**0.0110** / **0.0430** / **0.0176**
Wizard-of-Wikipedia (Test Seen)	Seq2Seq	0.1802 / 0.0608 / 0.0248	0.0480 / 0.2575 / 0.5468	0.0175 / 0.2427 / 0.0327
	PostKS	0.1936 / 0.0679 / 0.0278	0.0487 / 0.2642 / 0.5539	0.0245 / **0.3381** / 0.0457
	our model	**0.1961** / **0.0686** / **0.0281**	**0.0500** / **0.2759** / **0.5709**	**0.0248** / 0.3255 / **0.0461**
Wizard-of-Wikipedia (Test Unseen)	Seq2Seq	0.1735 / 0.0560 / 0.0227	**0.0397** / **0.2148** / 0.4849	0.0143 / 0.1827 / 0.0264
	PostKS	0.1805 / 0.0575 / 0.0221	0.0325 / 0.2077 / **0.4951**	0.0208 / 0.2557 / 0.0384
	our model	**0.1808** / **0.0585** / **0.0228**	0.0318 / 0.2001 / 0.4805	**0.0210** / **0.2620** / **0.0388**

Table 2. The ablation results on the PersonaChat dataset of our model.

Models	BLEU-1 / 2 / 3	Distinct-1 / 2 / 3	Knowledge-R / P / F1
our model	**0.1799** / **0.0739** / **0.0333**	0.0144 / **0.1011** / **0.2804**	**0.0110** / 0.0430 / **0.0176**
w/o prior contrast	0.1745 / 0.0725 / 0.0330	**0.0146** / 0.0995 / 0.2748	0.0102 / 0.0414 / 0.0163
w/o posterior contrast	0.1785 / 0.0737 / 0.0333	0.0144 / 0.0991 / 0.2739	0.0109 / **0.0437** / 0.0175

- **Seq2Seq** [4] is an attentive seq2seq model that does not have access to external knowledge.
- **PostKS** [10] is an attentive seq2seq that selects knowledge with the posterior distribution in the training phase and the prior distribution in the testing phase.

Implementation. We use a bidirectional GRU with 400 hidden states for each layer as our encoder and 1-layer GRUs with 800 hidden states in our decoder. All encoders and decoders do not share any parameters. We set the word embedding size to be 300 and initialized it using GloVe [13]. We use a vocabulary table that has no more than 20,000 words. We use an Adam optimizer [9], where the batch size of 16, and the learning rate is 5e-4. In the first five epochs, we minimize the BOW loss only for pre-training the knowledge selector. In the remaining epochs, we minimize the sum of all losses. We evaluated our model on the validation set every 100 steps and stopped training when the model did not update the minimal loss for a whole epoch.

Evaluation. We adopted several automatic metrics to perform the evaluation. *BLEU-1/2* and *Distinct-1/2* are two widely used metrics for evaluating the quality and diversity of generated responses. Due to the lack of labels, the quality of selected knowledge is hard to be measured directly. Following [10], we use *Kownledge-R/P/F1* to evaluate the knowledge quality of generated responses via measuring the relevancy between generated responses and the knowledge set. Specifically, given the set of non-stop words in a response Y and in the knowledge set K, denoted by W_Y and W_K, Knowledge-R/P/F1, denoted by $R/P/F1$ respectively, are defined as follows.

$$R = \frac{|W_Y \bigcap W_K|}{|W_K|} \tag{19}$$

$$P = \frac{|W_Y \bigcap W_K|}{|W_Y|} \tag{20}$$

$$F1 = 2 \cdot \frac{R \cdot P}{R + P} \tag{21}$$

4.2. Evaluation Results.

Effect of Dual Contrastive Learning The evaluation results are summarized in Table 1. Bold numbers show the best results among all models. We observe that our model outperforms baseline models in terms of knowledge utilization of generated responses on almost all datasets. For example, Knowledge-R/P/F1 on PersonaChat is increased from 0.0098/0.0407/0.0158 (PostKS) to 0.0110/0.0430/0.0176 (our model), indicating the improvement in terms of the quality of knowledge selection.

Ablation Study The ablation results on PersonaChat of our model are reported in Table 2. We observe that both prior and posterior contrast contribute to our model because the performance degrades without any of them. By comparison, the prior contrast contributes more, where we think the reason is that the prior distribution is directly used in the testing phase. The prior contrast directly increases knowledge selection compared to the posterior contrast.

5. Conclusion

This paper proposes a dual contrastive method for unsupervised knowledge selection in dialogue systems, which is the first work that introduces contrastive learning into knowledge selection in dialogue systems. Experiment results show that our model has improved on diversity and knowledge incorporation of generated responses. As for future work, we plan to extend our contrastive method to Transformer-based architecture.

ACKNOWLEDGMENTS

This research was supported in part by the Ministry of Science and Technology of China under Grant No. 2020AAA0104200, and the Shenzhen Key Laboratory of Marine IntelliSense and Computation under Contract ZDSYS20200811142605016.

References

[1] BAHDANAU, D., CHO, K., AND BENGIO, Y. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR* (2015).

[2] CAI, H., CHEN, H., SONG, Y., DING, Z., BAO, Y., YAN, W., AND ZHAO, X. Group-wise contrastive learning for neural dialogue generation. In *Proc. of EMNLP* (2020).

[3] CHEN, X., MENG, F., LI, P., CHEN, F., XU, S., XU, B., AND ZHOU, J. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proc. of EMNLP* (2020).

[4] CHO, K., VAN MERRIENBOER, B., BAHDANAU, D., AND BENGIO, Y. On the properties of neural machine translation encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014* (2014).

[5] DINAN, E., ROLLER, S., SHUSTER, K., FAN, A., AULI, M., AND WESTON, J. Wizard of wikipedia knowledge-powered conversational agents. In *Proc. of ICLR* (2019).

[6] GHAZVININEJAD, M., BROCKETT, C., CHANG, M., DOLAN, B., GAO, J., YIH, W., AND GALLEY, M. A knowledge-grounded neural conversation model. In *Proc. of AAAI* (2018).

[7] GUTMANN, M., AND HYVARINEN, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.* (2012).

[8] KIM, B., AHN, J., AND KIM, G. Sequential latent knowledge selection for knowledge-grounded dialogue. In *Proc. of ICLR* (2020).

[9] KINGMA, D. P., AND BA, J. Adam A method for stochastic optimization. In *Proc. of ICLR* (2015).

[10] LIAN, R., XIE, M., WANG, F., PENG, J., AND WU, H. Learning to select knowledge for response generation in dialog systems. In *Proc. of IJCAI* (2019).

[11] LIN, X., JIAN, W., HE, J., WANG, T., AND CHU, W. Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In *Proc. of ACL* (2020).

[12] PARTHASARATHI, P., AND PINEAU, J. Extending neural generative conversational model using external knowledge sources. In *Proc. of EMNLP* (2018).

[13] PENNINGTON, J., SOCHER, R., AND MANNING, C. D. Glove global vectors for word representation. In *Proc. of EMNLP* (2014).

[14] REN, P., CHEN, Z., MONZ, C., MA, J., AND DE RIJKE, M. Thinking globally, acting locally distantly supervised global-to-local knowledge selection for background based conversation. In *Proc. of AAAI* (2020).

[15] SHANG, L., LU, Z., AND LI, H. Neural responding machine for short-text conversation. In *Proc. of ACL* (2015).

[16] TIAN, Z., BI, W., LEE, D., XUE, L., SONG, Y., LIU, X., AND ZHANG, N. L. Response-anticipated memory for on-demand knowledge integration in response generation. In *Proc. of ACL* (2020).

[17] VINYALS, O., AND LE, Q. V. A neural conversational model. *CoRR abs1506.05869* (2015).

[18] YAO, L., ZHANG, Y., FENG, Y., ZHAO, D., AND YAN, R. Towards implicit content-introducing for generative short-text conversation systems. In *Proc. of EMNLP* (2017).

[19] ZHANG, S., DINAN, E., URBANEK, J., SZLAM, A., KIELA, D., AND WESTON, J. Personalizing dialogue agents I have a dog, do you have pets too. In *Proc. of ACL* (2018).