Computer Speech & Language 63 (2020) 101069

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Knowledge-Grounded Response Generation with Deep Attentional Latent-Variable Model

Hao-Tong Ye, Kai-Lin Lo, Shang-Yu Su, Yun-Nung Chen Ph.D.*

National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan

ARTICLE INFO

Article History:
Received 30 July 2019
Revised 1 January 2020
Accepted 16 January 2020
Available online 13 February 2020

Keywords:
Knowledge-grounded
Response generation
Variational model

ABSTRACT

End-to-end dialogue generation has achieved promising results without using handcrafted features and attributes specific to each task and corpus. However, one of the fatal drawbacks in such approaches is that they are unable to generate informative utterances, so it limits their usage from some real-world conversational applications. In order to tackle this issue, this paper attempts to generate diverse and informative responses with a variational generation model, which contains a joint attention mechanism conditioning on the information from both dialogue contexts and extra knowledge. The experiments on benchmark DSTC7 data show that the proposed method generates responses with more grounded knowledge and improve the diversity of generated language.

1. Introduction

Dialogue-related research can be mainly categorized into two branches: (1) task-oriented dialogues: systems trying to help users complete a certain task (2) chit-chat: systems that can handle casual conversations that do not belong to any specific domain. Recently, how to bridge these two branches has become a new research direction in conversation modeling, where the system can generate useful and fact-grounded responses via external knowledge without domain constraints (D'Haro et al., 2020; Hori et al., 2019; Ghazvininejad et al., 2018).

Prior work showed that end-to-end neural models are capable of generating sound responses for chit-chat dialogues in a data-driven way, without using handcrafted features specific to each corpus or different task (Sordoni et al., 2015; Vinyals and Le, 2015; Gao et al., 2018; Li et al., 2016). However, such systems still highly rely on the information stored in training corpora, which is constrained by time, space, and speakers during data collection. Also, those systems lack direct access to external information and knowledge-grounded mechanism; therefore they cannot effectively retrieve real-world common senses and facts in order to respond properly. This fundamental limitation makes end-to-end systems difficult to complete tasks (Li et al., 2017; Peng et al., 2018) or generate fact-grounded chit-chat (Ghazvininejad et al., 2018).

On the other hand, for traditional dialogue systems, we can easily insert external knowledge and facts into models at the cost of hand-coding detailed features, which requires a large amount of pre-processing and data labeling. For those tasks or corpora related to complex information or professional knowledge, pre-processing and annotations are difficult to acquire, thus making this approach impractical.

In this work, we propose an end-to-end variational model with the attention mechanism that models the interaction between dialogue contexts and external knowledge. This model strikes a balance between *scalability* and *generalization* of neural models and provides more factual and knowledge-grounded responses compared to the traditional systems. Such extension is especially

*Corresponding author.
E-mail addresses: r08922065@csie.ntu.edu.tw (H.-T. Ye), b04902010@csie.ntu.edu.tw (K.-L. Lo), f05921117@ntu.edu.tw (S.-Y. Su), y.v.chen@ieee.org, y.v.chen@ieee.org (Y.-N. Chen).

https://doi.org/10.1016/j.csl.2020.101069
0885-2308/

Table 1
The example from subreddit *todayilearned*. The horizontal lines indicate the tree-structure of the conversation, where the last two responses share the same contexts. The shown fact retrieved by our model is considered the most relevant to the given conversation among all facts extracted by the official script from the Wikipedia page.

Conversation:
til monty python member terry gilliam was author j. k rowling 's first choice to direct the first harry potter movie, but was rejected for chris columbus. in an interview he said "i was the perfect guy to do harry potter ... i mean, chris columbus ' versions are terrible. just dull. pedestrian " https://en.wikipedia.org/wiki/terry_gilliam
—gilliam would have been great - but we 'd still be waiting on the second movie.
——he should do an animated version
———harry potter & the giant soft gradient foot
—they hired chris columbus due to his experience directing child actors.
—— i also think he 's really good at seeing things from a kid 's imagination. those first 2 movies really seemed like someone went into my head and said " ok we 're going to film a movie here!"—— came here to say this. iirc, he was hired specifically because he was good with kids... which gilliam had little experience with. i think they turned out very well, very true to the books.
—— came here to say this. iirc, he was hired specifically because he was good with kids... which gilliam had little experience with. i think they turned out very well, very true to the books.
Retrieved top-1 fact:
j. k. rowling, the author of the harry potter series, is a fan of gilliam's work. consequently, he was rowling's first choice to direct harry potter and the philosopher's stone in 2000, but warner bros. ultimately chose chris columbus for the job. [32] in response to this decision, gilliam said that "i was the perfect guy to do harry potter. i remember leaving the meeting, getting in my car, and driving for about two hours along mulholland drive just so angry. i mean, chris columbus ' versions are terrible. just dull. pedestrian. " [33] in 2006, gilliam said that he found alfonso cuarn ' s harry potter and the prisoner of azkaban to be " really good ... much closer to what i would've done. ' [34] in retrospect, however, gilliam has stated that he wouldn't have liked to direct any potter film. in a 2005 interview with total film, he said that he would not enjoy working on such an expensive project because of interference from studio executives. [35]

important for a conversational model deployed in systems that require more relevant and informative interactions (e.g. recommendation systems).

To test the ability of generating knowledge-grounded responses, the Seventh Dialog System Technology Challenge (DSTC7) proposed a benchmark Reddit dataset, in which the conversations are accompanied with a link to an external webpage that may contain related facts and knowledge. A dataset example is shown in Table 1, where the last two responses share the same contexts, and the fact retrieved by our model contains related knowledge given the conversation. Following the idea, more knowledge-grounded conversational data was collected and published for investigating this research direction in terms of diverse aspects (Moon et al., 2019; Gopalakrishnan et al., 2019).

2. Proposed approach

The task is to generate a suitable response that contains grounded knowledge or factual information given its conversational contexts.

2.1. Model framework

The main difference between this task and others is the inclusion of context-relevant facts, which are retrieved from website links mentioned at the beginning of the conversation. This external knowledge provides our model cues about how to infuse responses with more information. Therefore, we first build a retrieval model to effectively obtain facts containing relevant knowledge and then learn the conversation model to generate knowledge-grounded responses. Below we describe the detail of the proposed conversation model, where given conversation contexts and its related facts, the goal is to generate an informative response.

2.2. Conversation model

For each conversation, our model takes as input the dialogue context $\mathbf{C}$ and context-relevant facts $\mathbf{F}$, and outputs the fact-grounded response $\mathbf{R}$. Specifically, $\mathbf{C}=\{c_i\}_{i=1}^{N_c}$, where $c_n=\{c_{nj}\}_{j=1}^{T_n^c}$ is a sequence of word embeddings in the n-th utterance of the conversation. For the fact, $\mathbf{F}=\{f_i\}_{i=1}^{N_f}$, where $f_n=\{f_{nj}\}_{j=1}^{T_n^f}$ is a sequence of word embeddings of the n-th fact. The generated response is formulated as $\mathbf{R}=\{r_i\}_{i=1}^{T^r}$. In our model, we treat the conversation utterances and facts as two sequences, with a special token used to separate individual utterances or facts; that is, the contexts and facts are turned into $\mathbf{C}=\{c_i\}_{i=1}^{N}$ and $\mathbf{F}=\{f_j\}_{j=1}^{M}$ respectively, where c_i and f_j are word embedding vectors.

First, we use two separate encoders, Enc_C and Enc_H, to encode the dialogue contexts and facts respectively. The encoded contexts and facts $H_C=\{h_i^c\}_{i=1}^{N}$, $H_F=\{h_j^f\}_{j=1}^{M}$ are fed into the attention module, and then the decoder generates the fact-grounded response. In our model, with the encoded contexts and facts, the decoder generates the response in an auto-regressive way, which is commonly called as a sequence-to-sequence model. For each step, the output of the decoder o_t is calculated from previous output o_{t-1} and the encoded information H_C and H_F:

$$o_t=\mathrm{Dec}\big(o_{t-1},\mathrm{Attn}(o_{t-1},H_C,H_F)\big). \tag{1}$$

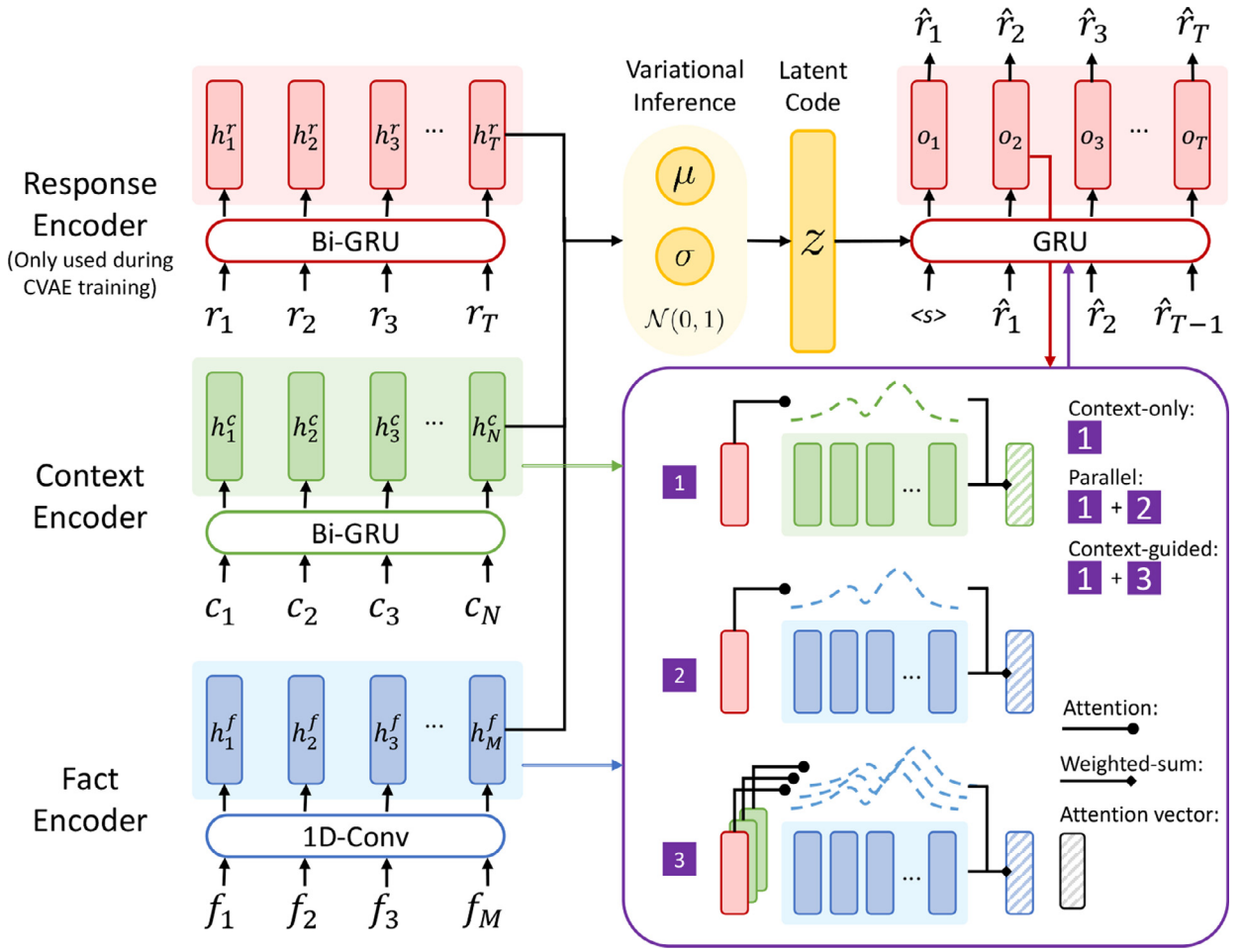

Fig. 1. Illustration of the proposed model architecture.

The output of the decoder, o_t, is then projected to the vocabulary through a linear layer followed by a softmax activation. The proposed model is illustrated in Fig. 1, where several encoders focus on different types of information. During generation, this model proposes 1) a *fact-grounded attention* mechanism that can explicitly consider the contexts and facts and 2) a *conditional variational generation* model that can produce diverse and informative responses. The detail of the two modules is described below.

2.3. Fact-Grounded attention

Unlike previous work that generated responses given only a single source of the input document (Shao et al., 2017; Mei et al., 2017; Xing et al., 2018), our model draws relevant information from related 'facts'. In order to capture the relations between these three types of information, dialogue contexts, facts, and responses, we apply three attention variants to model their interactions (Bahdanau et al., 2015): *context-only attention, parallel attention*, and *context-guided fact attention* detailed below.

2.3.1. Context-Only attention

One simple attention baseline only uses the information from contexts to generate the response. That is, with the last-step output o_{t-1} and the encoded information H_C, H_F, the attention is calculated as:

$$\begin{aligned} \text{Attn}(o_{t-1}, H_C, H_F) &= \sum_{i=1}^{N} \alpha_i^t h_i^c, \\ e_i^t &= v_c^T \tanh(W_1^c o_{t-1} + W_2^c h_i^c) + b_c, \\ \alpha^t &= \text{softmax}(e^t), \end{aligned} \tag{2}$$

where v_c, W_1^c, W_2^c, and b_c are trainable parameters.

2.3.2. Parallel attention

In order to well utilize the facts, one trivial solution is to consider facts as the additional contexts; that is:

$$\begin{aligned} \text{Attn}(o_{t-1}, H_C, H_F) &= \left[\sum_{i=1}^{N} \alpha_i^t h_i^c; \sum_{j=1}^{M} \beta_j^t h_j^f\right] \\ m_j^t &= v_f^T \tanh(W_1^f o_{t-1} + W_2^f h_j^f) + b_f, \\ \beta^t &= \text{softmax}(m^t), \end{aligned} \tag{3}$$

and α^t is the same as the context-only attention. v_f, W_1^f, W_2^f, b_f are trainable parameters.

2.3.3. Context-Guided fact attention

To better model the interaction between contexts and facts, we propose to use the information from contexts to guide the attention towards facts. Specifically, we modify the attention on facts from the parallel attention as shown below. We first calculate the attention distribution from contexts to facts,

$$\begin{aligned} M_{i,j} &= v_g^T \tanh(W_1^g h_i^c + W_2^g h_j^f) + b_g, \\ m_j^c &= \sum_{i=1}^{N} M_{i,j}, \\ \beta^c &= \text{softmax}(m^c). \end{aligned} \tag{4}$$

Then, for each step, we calculate the attention from the last-step output to facts, and take the mean of two distributions as the final attention distribution on facts:

$$\begin{aligned} \widehat{m}_i^t &= v_o^T \tanh(W_1^o o_{t-1} + W_2^o h_i^c) + b_o, \\ \widehat{\beta}^t &= \text{softmax}(\widehat{m}^t), \\ \beta^t &= \frac{\widehat{\beta}^t + \beta^c}{2}, \end{aligned} \tag{5}$$

where v_g, v_o, W_1^g, W_2^g, W_1^o, W_2^o, b_g, and b_o are trainable parameters. Hence, the obtained attention is guided by the contextual information.

2.4. Conditional variational generation

The conversations in the dataset for the DSTC7 challenge are tree-like structures, where for each context, there may be multiple reference responses. This is also an important perspective for the natural conversations: for arbitrary dialogue contexts, there are usually various ways to respond to it.

With the above consideration, we take the benefit from the variational autoencoder for tackling this task (Bahuleyan et al., 2018; Du et al., 2018; Le et al., 2018; Serban et al., 2017; Shen et al., 2018; Zhao et al., 2017; Gu et al., 2019), because this model has the better capability of capturing such relation than a simple seq2seq model according to the prior studies (Sohn et al., 2015; Diederik and Welling, 2014). To our best knowledge, this work is the first attempt that utilizes the variational model for generating knowledge-grounded conversations. Note that Ruan et al. (2019) proposed to use variational model for this task at the same time. The detail of the proposed variational model is described below.

2.4.1. CVAE For dialogue generation

For each conversation, we represent it via four random variables: the desired response R, the contexts and facts, C and F, and a latent variable z. The conditional probability $p(R, z|C, F)$ can be rewritten as:

$$p(R, z|C, F) = p(R|C, F, z)p(z|C, F). \tag{6}$$

We model the probability $p(R|C, F, z)$ and $p(z|C, F)$ using the parameters θ and ϕ respectively. Under the variational autoencoder (VAE) framework, we can interpret θ and ϕ as the decoder and the encoder; by setting up a Bayesian prior $p(z|C, F)$, our optimization target $p_\theta(R|C, F)$ becomes the variational lower bound (ELBO):

$$\log p_\theta(R|C, F) \geq -\text{KL}\Big(q_\phi(z|R, C, F) \parallel p(z|C, F)\Big) + \mathbb{E}_{q_\phi(z|R,C,F)}[\log p_\theta(R|C, F, z)]. \tag{7}$$

In our model, the prior $p(z|C, F)$ is set as $\mathcal{N}(0, I)$.

2.4.2. Annealing loss of KL divergence

As mentioned above, the optimization target, which is the variational lower bound of $\log p_\theta(R|C, F)$, is composed of two subgoals: one is to minimize the KL divergence between the prior and the conditional encoder probability q_ϕ; another is to maximize the reconstruction probability.

Table 2
Statistics of the used dataset.

	Time Period	Before Filter	After Filter
Train	2015-01 ~ 2016-12	1,101,684	142,750
Dev	2017-01 ~ 2017-06	116,858	14,875

It is found that the model tends to minimize the KL divergence instead of reducing the reconstruction error during early training, resulting in a KL vanishing issue. In order to alleviate the strong bias on minimization of KL divergence, we apply the annealing loss trick to scale down the effect of the KL term at the beginning of training for improving the performance (Bowman et al., 2016).

2.5. Training

The proposed model is trained to generate the responses using the CVAE objective, where the attention mechanisms enforce the responses to cover the fact-related information for *knowledge-grounded response generation*.

3. Experiments

To evaluate the proposed model, we conduct the experiments on the DSTC7 challenge. The used dataset and the experimental setting are described below. Then the results are analyzed in terms of objective and subjective evaluation metrics.

3.1. Dataset

The dataset used in DSTC7-Track2 is crawled from Reddit with the scripts[1], which consists of discussions from subreddits like *todayilearned, worldnews, movies*, etc. In the dataset, the posts include a link to an external webpage, from which the facts for each conversation are then extracted.

In order to encourage our conversation model to contain factual information, we process this dataset to make sure the conversations in which the context and provided facts are relevant. The processing procedure is described as:

1. **Fact relevance**: Because the facts are extracted from the HTML source codes of webpages, some of them lack the relevant information (e.g. metadata), we use TF-IDF to rank all facts and keep the top-1 fact as the relevant knowledge for ensuring better data quality.
2. **Knowledge-grounded response**: Because the discussions in some conversations may deviate from the original topic, making all facts being irrelevant to the dialogue contexts, we thus filter out data samples where the response and the retrieved fact have no common words without considering punctuations and stopwords[2] This procedure ensures the training data to match our goal about knowledge-grounded responses.

Due to the limitation of computation resources (one GTX 1080), we use only a subset of training data collected from more recent posts (those within time period 2015-01 to 2016-12, which consist of half of the training data before filtering), and discard the data samples with the responses longer than 20. Table 2 shows the detailed statistics of the dataset after our processing.

3.2. Training details

Considering that the dataset contains a large number of Internet slangs and spoken English, we train a 100 dimension word embeddings via *GLoVe* from train and development conversations and facts (Pennington et al., 2014). We truncate the context to the last 100 tokens and the fact to the first 500 tokens.

The context encoder Enc_C is a 2-layer bidirectional GRU (Cho et al., 2014) with hidden size 128; the fact encoder Enc_H is a convolutional network with 1,2,3 width filters, and 128 feature maps per filter. The decoder *Dec* is a 2-layer unidirectional GRU with the hidden size 128. For the CVAE variants, another 2-layer bidirectional GRU with the hidden size 128 is used to encode the responses. We performed grid search for hyperparameters by training the CO model for 1 epoch and selected the combination with the best performance on dev set. Search space for context encoder and decoder {1, 2, 3}-layer and hidden size {128, 256, 512}. As for the fact encoder, since facts tend to be really long, CNN is chosen for its speed and we used similar hyperparameters from (Jacovi et al., 2018).

Our models are trained using the teacher-forcing mechanism to maximize the likelihood of generating $\mathbf{R} = \{r_i\}_{i=1}^{T^r}$. We used adam (Kingma and Ba, 2014) with the default setting as our optimizer. During testing, we apply beam search where the beam size is 8.

[1] https://github.com/DSTC-MSR-NLP/DSTC7-End-to-End-Conversation-Modeling
[2] We used stopwords defined in spaCy.

Table 3
The automatic evaluation of baselines and the proposed methods. The baseline is a context-to-response seq2seq model without attention. CO, PA, CG correspond to context-only attention, parallel attention and context-guided attention respectively.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Nist-1	Nist-2	Nist-3	Nist-4	METEOR
Baseline	2.861	0.566	0.143	0.041	0.007	0.007	0.007	0.007	2.439
CO	2.634	0.513	0.130	0.038	0.006	0.006	0.006	0.006	2.417
+CVAE	2.690	0.527	0.145	0.042	0.007	0.007	0.007	0.007	2.371
PA	3.698	0.763	0.200	0.063	0.020	0.021	0.021	0.021	2.574
+CVAE	2.449	0.538	0.129	0.033	0.009	0.009	0.009	0.009	2.301
CG	2.142	0.443	0.124	0.040	0.004	0.004	0.004	0.004	2.258
+CVAE	**3.898**	**0.817**	**0.223**	**0.074**	**0.023**	**0.024**	**0.024**	**0.024**	**2.620**

Table 4
The automatic evaluation of baselines and the proposed methods in terms of diversity and entropy scores.

Model	Diversity-1	Diversity-2	Entropy-1	Entropy-2	Entropy-3	Entropy-4
Baseline	0.004	0.012	3.937	4.958	5.504	5.996
CO	**0.013**	0.028	4.137	5.392	6.148	6.734
+CVAE	**0.013**	**0.030**	4.204	**5.436**	**6.239**	**7.049**
PA	0.012	0.027	**4.244**	5.378	6.040	6.576
+CVAE	0.011	0.027	4.120	5.338	6.125	6.775
CG	0.011	0.026	4.131	5.300	6.082	6.820
+CVAE	0.012	0.027	4.089	5.220	5.916	6.427

3.3. Results

In the experiments, we perform two sets of evaluation, automatic evaluation, and human evaluation, to better validate our generated results.

3.3.1. Automatic evaluation

Our evaluation metrics include BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), NIST (Doddington, 2002), diversity (Li et al., 2016a) and entropy (Zhang et al., 2018) scores, where the first three types focus on relevance and the last two types focus on diversity. We use the implementation in the Python package nlg-eval[3] for BLEU and METEOR scores (Sharma, El Asri, Schulz, Zumer, 2017), and the NLTK toolkit to calculate NIST scores. Our results are shown in Table 3 and Table 4.

It can be found that the context-guided attention model with CVAE (CG+CVAE) achieves better performance for most metrics in terms of the similarity between the generated responses and the ground truth responses. This justifies the effectiveness of our context-guided attention, because its goal is to generate responses containing more relevant knowledge, and the metrics slightly measure the relatedness. However, the context-only attention with CVAE (CO+CVAE) obtains the higher diversity, which is also important for this generation task. The results show the small improvement achieved by the proposed CVAE model in terms of the generation quality and diversity.

3.3.2. Human evaluation

In order to understand the effect of our fact-grounded attention and variational generation, we conduct human evaluation on three proposed methods: the parallel attention model as our baseline (PA), compared with the parallel attention with variational generation (PA+CVAE), and the context-guided attention (CG). First, we randomly sample 100 testing samples that fulfill the following two conditions:

1. Each response has at least 3 words because some methods tend to produce very short responses, which is hard to evaluate.
2. Due to the goal of fact-grounded generation, we make sure that the contexts and the retrieved fact have more than 3 common words for each sample, where punctuations and stop-words are not considered.

Then we conduct human evaluation for our proposed methods in a similar way to the official evaluation:

1. In addition to *relevance* and *interest*, which are asked in official evaluation, we ask the judges to evaluate two additional metrics: *fluency* and *knowledge relatedness* (to the retrieved fact) of our response.
2. Because we only pick one fact based on the contexts as our model input, we directly provide this fact to judges as the extra information for them to better evaluate *knowledge relatedness* of the response.

[3] https://github.com/Maluuba/nlg-eval

Table 5
Human evaluation results in our offline and the official evaluation.

	Model	Context Relevance	Interest	Fluency	Knowledge Relatedness	Average
Offline	PA	2.47 ± 0.86	2.37 ± 0.75	4.13 ± 0.85	2.19 ± 0.87	2.79
	PA+CVAE	2.40 ± 0.81	2.38 ± 0.77	4.00 ± 0.92	2.10 ± 0.86	2.72
	CG	2.25 ± 0.83	2.18 ± 0.76	3.86 ± 1.07	2.02 ± 0.83	2.58
Official	Submitted (CG+CVAE)	2.52 ± 0.04	2.40 ± 0.05	-	-	2.46
	Baseline	2.91 ± 0.05	2.68 ± 0.04	-	-	2.80
	Best	3.09 ± 0.04	2.87 ± 0.05	-	-	2.94
	Human	3.61 ± 0.04	3.49 ± 0.04	-	-	3.55

Table 6
Model response sample.

Retrieved top-1 fact:
in the united states, centenarians traditionally receive a letter from the president, congratulating them for their longevity. nbc ' s today show has also named new centenarians on air since 1983. centenarians born in ireland receive a 2540 'centenarians' bounty " and a letter from the president of ireland, even if they are resident abroad. [63] japanese centenarians receive a silver cup and a certificate from the prime minister of japan upon their 100th birthday, honouring them for their longevity and prosperity in their lives. swedish centenarians receive a telegram from the king and queen of sweden. [64] centenarians born in italy receive a letter from the president of italy. in japan, a " national respect for the aged day " has been celebrated every september since 1966.
Conversation:
– til in the united states, people who turn 100 years old receive a letter from the president, congratulating them on their longevity.
– same in canada but 90 instead of 100
Ground Truth: is that the canadian exchange rate these days ?
PA Response: they are the same thing.
PA+CVAE Response: you can have to be a.
CG Response: it's not the same thing in the uk.

In both offline and official settings, judges were asked to select a score from scale 1 to 5 (representing *strongly disagree, disagree, neutral, agree, strongly agree*). The results are shown in Table 5. The official baseline is a sequence-to-sequence RNN model that takes only context history as input. The submitted system, the best-achieved results, and human performance are also included in Table 5 for better comparison. Note that the numbers for two sets of evaluation may not be directly compared but for reference.

In the offline human evaluation, it is found that the proposed models do not achieve better performance and the difference between all models is small. From the official evaluation, our submitted results are also between *disagree* (2) and *neutral* (3) as in our evaluation, but the context-guided attention achieves slightly better numbers than other proposed models shown in the offline setting. Furthermore, the best-achieved performance is about 2.94, which is also lower than *neutral* (3), implying the difficulty of this task. It is clear that there is a huge gap between the currently machine-achieved and human-achieved performance, so this task requires further investigation.

3.4. Qualitative analysis

The above results tell that there is no significant difference between our proposed models and baselines. Some model response samples from the human evaluation set are shown in Table 6 for our qualitative analysis. In this example, adding CVAE generates a more diverse response than the parallel attention result, but may not effectively ground the knowledge in the sentence. Also, our context-guided result seems to focus more on the fact compared to other models. However, the ground truth in the data is very difficult to simulate for the current models, because it may need additional knowledge or common sense. From the current results achieved by our model, we conclude that this task still needs further investigation.

4. Conclusions

We describe a variational knowledge-grounded conversation system, which attempts at modeling the relations between dialogue contexts and external facts in an end-to-end fashion. It guides a potential research direction about how external information interacts with dialogues and how the machine can capture such interaction for better knowledge-grounded response generation. In the experiments on DSTC7, the results demonstrate the difficulty of this task, because almost all current models fail to generate reasonable responses. Therefore, the knowledge-grounded dialogue modeling requires further study in order to advance the machine's capacity of producing an informative and knowledgable conversation.

Acknowledgements

This work was financially supported from the Young Scholar Fellowship Program by Ministry of Science and Technology (MOST) in Taiwan, under Grant 108- 2636-E-002-003 and 108-2634-F-002-019.

References

Bahuleyan, H., Mou, L., Vechtomova, O., Poupart, P., 2018. Variational attention for sequence-to-sequence models. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1672–1682.
Banerjee, S., Lavie, A., 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65–72.
Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., Bengio, S., 2016. Generating sentences from a continuous space. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pp. 10–21.
D'Haro, L.F., Yoshino, K., Hori, C., Marks, T.K., Polymenakos, L., Kummerfeld, J.K., Galley, M., Gao, X., 2020. Overview of the seventh Dialog System Technology Challenge: DSTC7. Computer Speech & Language 62.
Doddington, G., 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the second international conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., pp. 138–145.
Du, J., Li, W., He, Y., Xu, R., Bing, L., Wang, X., 2018. Variational autoregressive decoder for neural response generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3154–3163.
Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., 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. pp. 1724-1734.
Ghazvininejad, M., Brockett, C., Chang, M.-W., Dolan, B., Gao, J., Yih, W.-t., Galley, M., 2018. A knowledge-grounded neural conversation model. Thirty-Second AAAI Conference on Artificial Intelligence.
Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., Hakkani-Tür, D., AI, A.A., 2019. Topical-chat: towards knowledge-grounded open-domain conversations. Proc. Interspeech 2019 1891–1895.
Ghazvininejad, M., Brockett, C., Chang, M.W., Dolan, B., Gao, J., Yih, W.T. and Galley, M., 2018. A knowledge-grounded neural conversation model. In: Thirty-Second AAAI Conference on Artificial Intelligence.
Hori, C., Perez, J., Higashinaka, R., Hori, T., Boureau, Y.L., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K., Kim, S., 2019. Overview of the sixth dialog system technology challenge. Dstc6. Computer Speech & Language 55, 1–25.
Jacovi, A., Shalom, O.S., Goldberg, Y., 2018. Understanding convolutional neural networks for text classification. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 56–65.
Gao, J., Galley, M. and Li, L., 2018. Neural approaches to conversational AI. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. pp. 2-7.
Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv:1412.6980
Le, H., Tran, T., Nguyen, T., Venkatesh, S., 2018. Variational memory encoder-decoder. Advances in Neural Information Processing Systems, pp. 1508–1518.
Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B., 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, pp. 110–119.
Diederik, P.K. and Welling, M., 2014. Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations.
Li, X., Chen, Y.-N., Li, L., Gao, J., Celikyilmaz, A., 2017. End-to-end task-completion neural dialogue systems. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 1, pp. 733–743.
Mei, H., Bansal, M., Walter, M.R., 2017. Coherent dialogue with attention-based language models. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI Press, pp. 3252–3258.
Moon, S., Shah, P., Kumar, A., Subba, R., 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, pp. 845–854.
Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, pp. 311–318.
Peng, B., Li, X., Gao, J., Liu, J., Chen, Y.-N., Wong, K.-F., 2018. Adversarial advantage actor-critic model for task-completion dialogue policy learning. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6149–6153.
Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.
Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M. and Gao, J., 2016. Deep reinforcement learning for dialogue generation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1192-1202.
Ruan, Y.-P., Ling, Z.-H., Liu, Q., Gu, J.-C., Zhu, X., 2019. Promoting diversity for end-to-end conversation response generation. arXiv:1901.09444
Serban, I.V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., Bengio, Y., 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. Thirty-First AAAI Conference on Artificial Intelligence.
Shao, Y., Gouws, S., Britz, D., Goldie, A., Strope, B., Kurzweil, R., 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2210–2219.
Sharma, S., El Asri, L., Schulz, H., Zumer, J., 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. CoRR abs/1706.09799.
Shen, X., Su, H., Niu, S., Demberg, V., 2018. Improving variational encoder-decoders in dialogue generation. Thirty-Second AAAI Conference on Artificial Intelligence.
Sohn, K., Lee, H., Yan, X., 2015. Learning structured output representation using deep conditional generative models. Advances in Neural Information Processing Systems, pp. 3483–3491.
Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., Dolan, B., 2015. A neural network approach to context-sensitive generation of conversational responses. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 196–205.
Vinyals, O., Le, Q., 2015. A neural conversational model. ICML Deep Learning Workshop 2015.
Xing, C., Wu, Y., Wu, W., Huang, Y., Zhou, M., 2018. Hierarchical recurrent attention network for response generation. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence.
Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X., Brockett, C. and Dolan, B., 2018. Generating informative and diverse conversational responses via adversarial information maximization. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 1815-1825.
Bahdanau, D., Cho, K. and Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations.
Zhao, T., Zhao, R., Eskenazi, M., 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 654–664.

Gu, X., Cho, K., Ha, J.W. and Kim, S., 2019. Dialogwae: Multimodal response generation with conditional Wasserstein auto-encoder. In: 7th International Conference on Learning Representations.

Hao-Tong Ye received bachelor degree in Computer Science and Information Engineering from National Taiwan University, Taipei, Taiwan in 2019. His research has been focused on conversational question answering.

Kai-Lin Lo received bachelor degree in Computer Science and Information Engineering from National Taiwan University, Taipei, Taiwan in 2019. His research has been focused on natural language generation.

Shang-Yu Su is a current Ph.D. student in Department of Computer Science and Information Engineering at National Taiwan University, Taipei, Taiwan. He holds the bachelor degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan. His research has been focused on deep learning and dialogue systems.

Yun-Nung Chen received Ph.D. in Computer Science from at Carnegie Mellon University, PA. She has been an assistant professor in the Department of Computer Science and Information Engineering at National Taiwan University, Taipei, Taiwan. Her research interests include spoken dialogue understanding, speech summarization, information extraction, and machine learning. Dr. Chen received Google Faculty Research Award and NVIDIA Scientific Research Award and currently serves a member in Speech and Language Technical Committee in IEEE.