Context-Sensitive Generation of Open-Domain Conversational Responses

Wei-Nan Zhang[*], **Yiming Cui**[†], **Yifa Wang**[*], **Qingfu Zhu**[*], **Lingzhi Li**[*], **Lianqiang Zhou**[‡], **Ting Liu**[*]

[*]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China.
[†]Joint Laboratory of HIT and iFLYTEK (HFL), iFLYTEK Research, Beijing, China
[‡]Joint Laboratory of HIT and Tencent Corporation, Shenzhen, China
[*]{wnzhang, yfwang, qfzhu, lzli, tliu}@ir.hit.edu.cn
[†]ymcui@iflytek.com
[‡]tomcatzhou@tencent.com

Abstract

Despite the success of existing works on single-turn conversation generation, taking the coherence in consideration, human conversing is actually a context-sensitive process. Inspired by the existing studies, this paper proposed the static and dynamic attention based approaches for context-sensitive generation of open-domain conversational responses. Experimental results on two public datasets show that the proposed static attention based approach outperforms all the baselines on automatic and human evaluation.

1 Introduction

Until recently, training open-domain conversational systems that can imitate the way of human conversing is still not a well-solved problem and non-trivial task. Previous efforts focus on generating open-domain conversational responses as an unsupervised clustering process (Ritter et al., 2010), a phrase-based statistical machine translation task (Ritter et al., 2011) and a search problem based on the vector space model (Banchs and Li, 2012), etc. With the booming of deep learning, particularly the neural network based sequence-to-sequence models, generating open-domain conversational responses gradually turns into an end-to-end encoding and decoding process (Sutskever et al., 2014; Vinyals and Le, 2015; Shang et al., 2015; Serban et al., 2016b; Li et al., 2016a; Li et al., 2016b; Shao et al., 2017; Yao et al., 2017). Despite the success of the above research on single-turn conversational response generation, human conversations are usually coherent (Li et al., 2016c) and **context-sensitive** (Tian et al., 2017; X-ing et al., 2017). Table 1 illustrates how contextual information in conversations impact on the response generation. For instance, given a message[1] "*How should I tell my mom?*", as input, to a single-turn conversational response generation model, it should output a fixed response regardless of the content in previous utterances. However, as shown in Table 1, in the conversations[2], the responses to be generated (the last utterance in Table 1) should not only dependent on the last one message ("*How should I tell my mom?*"), but also need to consider the longer **historical utterances** in the conversations.

Conversation 1	**Conversation 2**
A: I got a high score on my exam.	A: I failed to pass the exam.
B: Oh! Great!	B: That's too bad.
A: *How should I tell my mom?*	A: *How should I tell my mom?*
B: ***Go and give her a big surprise!***	B: ***Just tell her the truth and do well next time.***

Table 1: An example of the impact of contextual information on human conversations. "A" and "B" denote two speakers in the conversations.

[1] Here, a "message" indicates an input of a response in single-turn conversational response generation.

[2] In this paper, a "conversation" equals to an "open-domain conversation"and a "conversational response" or "response" equals to an "open-domain conversational response".

Proceedings of the 27th International Conference on Computational Linguistics, pages 2437–2447
Santa Fe, New Mexico, USA, August 20-26, 2018.

Recent studies on generating open-domain conversational responses begin to explore the context information to generate more informative and coherent responses. Serban et al. (2016a) presented a hierarchical recurrent encoder-decoder (HRED) to recurrently model the dialogue context. Serban et al. (2017b) further introduced a stochastic latent variable at each dialogue turn to improve the diversity of the HRED model. Zhao et al. (2017) proposed a conditional variational autoencoder based approach to learning contextual diversity for neural response generation. Xing et al. (2017) proposed a hierarchical recurrent attention network (HRAN) to jointly model the importance of tokens and utterances. Tian et al. (2017) treated the hierarchical modeling of contextual information as a recurrent process in encoding. We could make two conclusions from these works.

- First, existing studies of utterance modeling mainly focus on representing utterances by using bidirectional GRU (Xing et al., 2017) or unidirectional GRU (Tian et al., 2017).

- Second, there are two types of approaches on context (inter-utterance) modeling. One is the attention-based approach (Xing et al., 2017), the other is the sequential integration approach (Tian et al., 2017).

Drawing the advantages of the existing approaches, in this paper, we proposed a novel context-sensitive generation approach, which obtains the context representation of a conversation by weighing the importance of each utterance using two attention mechanisms, namely dynamic and static attention, to generate open-domain conversational responses.

2 The Proposed Context-Sensitive Generation Approach

2.1 Preliminary

A typical neural network based sequence-to-sequence model for generating open-domain conversational responses usually includes an encoder and a decoder. The encoder expresses an input message as a dense vector which represents the semantics of the input message. The decoder then generates a conversational response according to the semantic representation of the input message. In context-sensitive generation of open-domain conversational responses, the input message to the encoder usually includes several historical utterances in a conversation. Therefore, one of the key problems in context-sensitive generation is how to encode historical utterances in a conversation. Figure 1 presents two state-of-the-art approaches to encoding contextual information for context-sensitive response generation. Here, u_i, u_{i+1} and u_j denote the i-th, i+1-th and j-th utterance, respectively, in a conversation. As the inputs of the two models, they are then represented to utterance-level vectors as shown in the second layer of the two models in Figure 1. The context vectors of the two models are obtained by hierarchically representing the utterances to a dense vector c for decoding. It is easy to recognize that the frameworks used to illustrate the encoders of two existing context-sensitive generation models look similar to each other. There are two different parts between the two frameworks:

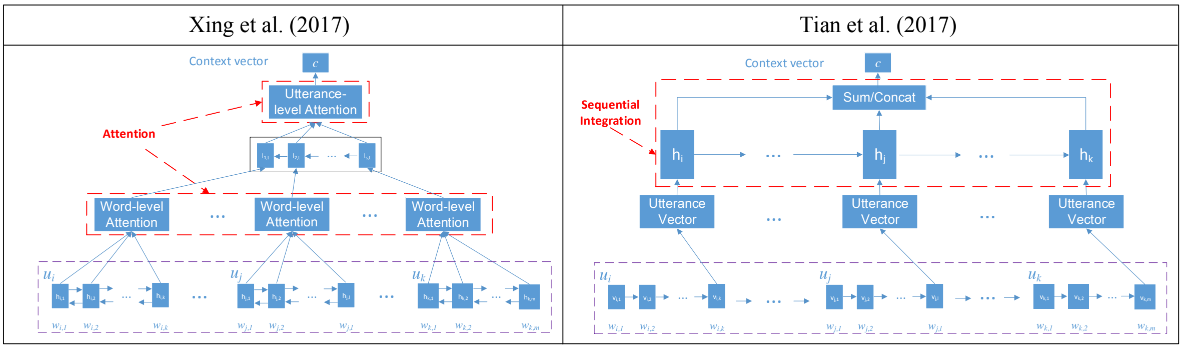


Figure 1: The encoders of two state-of-the-art approaches of open-domain conversational response generation using contextual information.

- **Utterance Representations: Bidirectional GRU vs. Unidirectional GRU**

Xing et al. (2017) utilized a bidirectional GRU and a word-level attention mechanism to transfer word representations to utterance representations. Tian et al. (2017) represented the utterance in a simpler way, which is a unidirectional GRU.

- **Inter-utterance Representations: Attention vs. Sequential Integration**

Xing et al. (2017) proposed a hierarchical attention mechanism to feed the utterance representations to a backward RNN to obtain contextual representation. Tian et al. (2017) proposed a weighted sequential integration (WSI) approach to use an RNN model and a heuristic weighting mechanism to obtain inter-utterance representation.

2.2 The Proposed Model

The proposed context-sensitive generation model is under the framework of encoder-decoder. To obtain the contextual representations, the proposed model consists of a hierarchical representation mechanism for encoding. For utterance representation, we consider the advantages of the two state-of-the-art approaches to encoding contextual information for context-sensitive response generation (Xing et al., 2017; Tian et al., 2017). We utilize a GRU model to obtain utterance representation. For inter-utterance representation, inspired by the above approaches of modeling inter-utterance representations, we proposed two attention mechanisms, namely dynamic and static attention, to weigh the importance of each utterance in a conversation and obtain the contextual representation. Figure 2 shows the framework of the proposed context-sensitive generation model. Drawing the advantages of attention mechanism on weighing the importance of utterances for generating open-domain conversational responses (Xing et al., 2017), we thus model the inter-utterance representation to obtain the context vector in two measures, namely static and dynamic attention, as shown in Figure 2. We then formally describe the static and dynamic attention for decoding process.

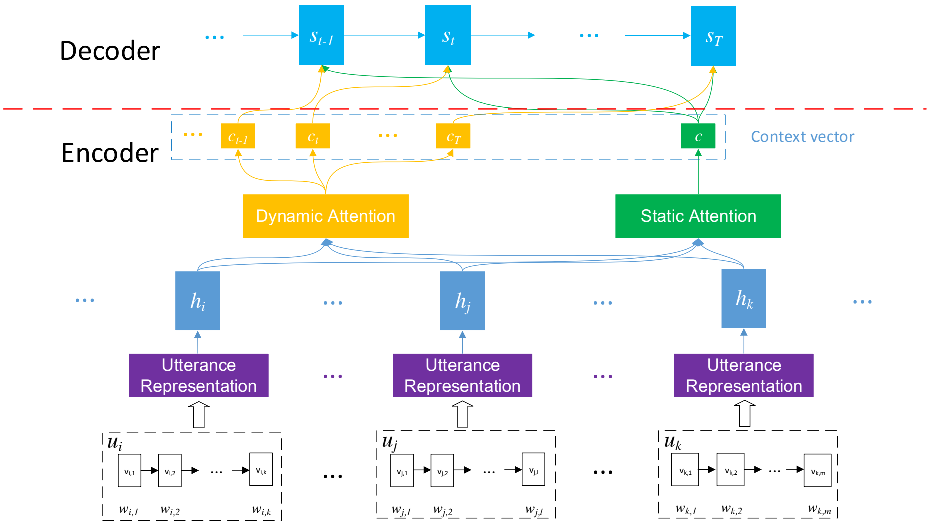


Figure 2: The proposed context-sensitive generation model for open-domain conversational response. Here, u_* denotes the $*$-th utterance in a conversation.

- **Static Attention based Decoding**

As shown in Figure 2, the static attention mechanism calculates the importance of each utterance as e_i or α_i $(i \in \{1, ..., s\})$.

$$e_i = V^T \tanh(Wh_i + Uh_s) \tag{1}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_i \exp(e_i)} \tag{2}$$

$$c = \sum_i \alpha_i h_i \tag{3}$$

Here, h_i and h_s denote the representations of hidden state of the i-th and the last utterance in a conversation, respectively. V, W and U are parameters. We can see that once the weights of each utterance α_i ($i \in \{1, ..., s\}$) are produced, they will be unchanged in the decoding process. In decoding, the t-th hidden state s_t can be calculated as follows:

$$s_t = f(y_{t-1}, s_{t-1}, c) \tag{4}$$

Here, y_{t-1} is the $t-1$-th output of the decoder and s_{t-1} is the hidden state of t-1-th time step in decoding. Notice that y_0 is set to be a special character and s_0 is initialized by h_s. The generated response is thus represented as a sequence of $y_1, y_2, ..., y_T$, where T denotes the last time step.

- **Dynamic Attention based Decoding**

Rather than the static attention mechanism fixes the weights of each utterance before decoding process, the dynamic attention mechanism maintains a weighting matrix and updates the weights of each utterance during decoding process as shown in Figure 2. The formal illustration of the dynamic attention mechanism for decoding is as follows:

$$e_{i,t} = V^T \text{tanh}(Wh_i + Us_{t-1}) \tag{5}$$

$$\alpha_{i,t} = \frac{\text{exp}(e_{i,t})}{\sum_i \text{exp}(e_i)} \tag{6}$$

$$c_t = \sum_i \alpha_{i,t} h_i \tag{7}$$

Here, V, W and U are also parameters that are independent to those in the static attention. T denotes the transposition operation of V. The $e_{i,t}$ and $\alpha_{i,t}$ are calculated in each time step t of decoding. The t-th hidden state s_t in dynamic attention-based decoder can be calculated as follows:

$$s_t = f(y_{t-1}, s_{t-1}, c_t) \tag{8}$$

The main difference between our proposed conversational response generation model and the above two state-of-the-art models is the two attention mechanisms for obtaining the contextual representation of a conversation. Rather than use a hierarchical attention neural network (Xing et al., 2017) to obtain the contextual representation of a conversation, we propose two utterance-level attentions for weighting the importance of each utterance in the context, which is more simple in structure and has less number of parameters than the hierarchical attention approach. Meanwhile, rather than use a heuristic approach to weigh the importance of each utterance in the context (Tian et al., 2017), in our proposed approach, the weights of utterance in the context are learned by two attention mechanisms from the data, which is more reasonable and flexible than the heuristic based approach.

3 Experimental Results

3.1 Experiment Settings

Dataset: Two datasets are selected for the experiment of generation of open-domain conversational responses. First is the Ubuntu dataset which is developed by Lowe et al. (2015). The dataset is extracted from the Ubuntu Internet Relayed Chat (IRC) channel and recently used for the generation of conversational responses in (Serban et al., 2016a; Serban et al., 2017b; Serban et al., 2017a). We follow the train-test split proposed by Serban et al. (2017a). It is worthy to note that there is no development set in Serban et al. (2017a). In this paper, we randomly select the same number of sessions to that in the test set from the training set. Second is the OpenSubtitles dataset which is proposed by Tiedemann (2009) and also used by Li et al. (2016a; Li et al. (2016c). The detailed statistics of the two datasets are shown in Table 2. It is worthy to note that the original released data of OpenSubtitles consists of about 40,000,000 utterances without partitions of conversational session, which is called "session" for short in the following of this paper. Therefore, we split each of 10 continuous utterances as a session. We then randomly sample 800,000 sessions for training (including 8,000 sessions for developing) and remove them from the complete dataset. In the rest of the complete dataset, we again randomly sample 8,000 sessions for testset. The vocabulary size equals to the number of unique tokens in both two datasets, respectively.

	Ubuntu	OpenSubtitles
Train size	429,915	792,000
Dev size	18,920	8,000
Test size	18,920	8,000
Vocabulary size	155,490	91,405
Avg # of u per session	7.5	10
Avg # of w per u	12.3	7.5

Table 2: The statistics of two experimental datasets. Avg is short for average. # represents number. u and w denote utterance and word respectively. The unit of training and test is a conversational session.

Hyper-Parameters: For the static attention model, the dimension of hidden layer in encoder and decoder is 512. The padding length is set to 15. The dimension of word embedding equals to 200. The word embedding is pre-trained using the skip-gram model in word2vec (Mikolov et al., 2013) and fine-tuned during the learning process. For the dynamic attention model, the dimension of hidden layer in encoder and decoder is 1024. The padding length and dimension of word embedding are same to the static attention model. Adam is used for optimization. The initial learning rate is 0.001 and the weight_decay is set to 10^{-5}. The dropout parameter equals to 0.5. Mini-batch is used and the batch size equals to 80. The number of iterations in training is 10.

Baselines: For the experimental comparisons, six baselines are chosen. Four out of them are state-of-the-art approaches. They are VHRED, CVAE, WSI, and HRAN.

- **LSTM:** Under the sequence-to-sequence framework for generation of conversational responses, the most simple but natural idea is to directly use the LSTM to encode all the utterances in a session word by word and then decode to generate a response.
- **HRED:** The first hierarchical recurrent encoder-decoder model, which is proposed by Serban et al. (2016a), for conversational response generation.
- **VHRED:** The augmented HRED model, which incorporates a stochastic latent variable at utterance level for encoding and decoding, is proposed by Serban et al. (2017b).
- **CVAE:** The conditional variational autoencoder based approach, which is proposed by Zhao et al. (2017), to learn context diversity for conversational responses generation.
- **WSI** and **HRAN** are proposed by Tian et al. (2017) and Xing et al. (2017) respectively. We detailed describe and compare the two models in Section 2.1 and 2.2 and their frameworks are shown in Figure 1.

3.2 Evaluation and Results

3.2.1 Automatic Evaluation

Until now, automatically evaluating the quality of open-domain conversational response is still an open problem. The BLEU score (Papineni et al., 2002), which is a widely used evaluation metric for machine translation, is not a suitable metric for conversation generation, as the appropriate responses to the same message may share less common words. Moreover, it is also impossible to construct a reference set, which includes all appropriate responses, of each message. The perplexity that is used to evaluate language model, is also not suitable to evaluate the relevance between messages and responses (Shang

Models	Ubuntu			OpenSubtitles		
	Average	Greedy	Extrema	Average	Greedy	Extrema
LSTM	0.2300	0.1689	0.1574	0.5549	0.5029	0.3897
HRED	0.5770	0.4169	0.3914	0.5571	0.5033	0.3932
VHRED	0.5419	0.3839	0.3627	0.5248	0.4821	0.3556
CVAE	0.5672	0.3982	0.3689	0.4708	0.3390	0.3173
WSI	0.5775	0.4196	0.3893	0.5598	0.4964	0.3903
HRAN	0.5964	0.4139	0.3898	0.5617	0.5195	0.3898
Dynamic$_{\rightleftarrows}$	0.5750	0.4043	0.3802	0.5487	0.5054	0.3812
Dynamic$_{\rightarrow}$	0.5968	0.4132	0.3877	0.5629	0.5193	0.3905
Static$_{\rightleftarrows}$	0.5998	0.4124	0.3886	0.5475	0.5147	0.3862
Static$_{\rightarrow}$	**0.6121**†	**0.4293**†	**0.3975**†	**0.5656**†	**0.5232**†	**0.3937**†

Table 3: The results of automatic evaluation on Ubuntu and OpenSubtitles datasets. Dynamic and Static are our proposed approaches whose framework is shown in Figure 2. The other models are baselines. $\rightarrow$ and $\rightleftarrows$ denote the use of unidirectional and bidirectional GRU in the proposed model to obtain utterance representations, respectively. † denotes the results pass the statistical significance test with $p < 0.05$.

et al., 2015; Li et al., 2016c). In this paper, we employ an evaluation metric that is proposed by Serban et al. (2016a) and also used in (Serban et al., 2017b). Rather than calculating the token-level or n-gram similarity as the perplexity and BLEU (Papineni et al., 2002), the metric measure the semantic similarity between a generated responses $\hat{r}$ and the ground-truth responses r by matching their semantic representations. The metric also has three aspects, namely $\mathbf{Average}$, $\mathbf{Greedy}$ and $\mathbf{Extrema}$. For the $\mathbf{Average}$, it first calculates the element-wise arithmetic average of embeddings of all words in $\hat{r}_a$ and r_a, respectively and produces two response representations $v_{\hat{r}_a}$ and v_{r_a}. The value of $\mathbf{Average}$ is then equals to the cosine similarity of $v_{\hat{r}_a}$ and v_{r_a}. For the $\mathbf{Greedy}$, every word in $\hat{r}$ will find a most similar word in r by calculating the cosine similarity of their word embeddings. After that, the element-wise arithmetic average of embeddings of all words in $\hat{r}_a$ and the corresponding words in r are calculated and two response representations $v_{\hat{r}_g}$ and v_{r_g} are produced. The value of $\mathbf{Greedy}$ is then equals to the cosine similarity of $v_{\hat{r}_g}$ and v_{r_g}. For the $\mathbf{Extrema}$, two embedding matrices $m_{\hat{r}}$ and m_r can be obtained by arranging the embeddings of all words in $\hat{r}_a$ and r_a, respectively. The i-th column of $m_{\hat{r}}$ is the embedding of the i-th word in $\hat{r}$ as well as that in m_r. Getting the maximum value of each row in $m_{\hat{r}}$ and m_r, respectively, we then obtain two response representations $v_{\hat{r}_e}$ and v_{r_e}. The value of $\mathbf{Extrema}$ is then equals to the cosine similarity of $v_{\hat{r}_e}$ and v_{r_e}.

Table 3 shows the experimental results on two datasets.

We can see that our proposed context-sensitive generation model with static attention outperforms all the baselines in the two datasets. It verifies the effectiveness of the proposed utterance-level attention mechanism on modeling context representations for generating conversational responses. To compare the dynamic and static attentions, we find that for the generation of conversational response, dynamically estimate the importance of each utterance in context performs worse than the static attention approach. The reason may be that the context vector in dynamic attention model is changed in every time step of decoding. The change of context vector may lead to decoding incoherent responses. Meanwhile, the unidirectional GRU based models outperform the bidirectional GRU based models. It doesn't illustrate the unidirectional GRU is better than the bidirectional GRU in utterance representation. It only indicates that in the current experimental settings, the unidirectional GRU based model outperforms the bidirectional one.

3.2.2 Human Evaluation

For human evaluation, we proposed 2 metrics, namely **Coherence** and **Naturalness**. As the example shown in Table 1, in context-sensitive generation of conversational responses, a generated response should not only dependent on the last one message but also need to consider the longer context in the

Models	Ubuntu			OpenSubtitles		
	Coherence	Naturalness	Diversity	Coherence	Naturalness	Diversity
LSTM	0.930	0.477	0.069	0.963	0.443	0.099
HRED	0.967	0.490	0.141	0.963	0.443	0.098
VHRED	1.010	0.507	0.140	0.986	0.473	0.093
CVAE	0.987	**0.513**	0.140	1.000	0.477	**0.114**
WSI	1.010	0.507	0.141	1.013	0.490	0.110
HRAN	1.027	0.510	0.147	**1.033**	0.477	0.109
Dynamic	0.987	0.507	**0.158**	1.013	0.477	0.109
Static	**1.070**	**0.513**	0.150	1.027	**0.497**	0.110

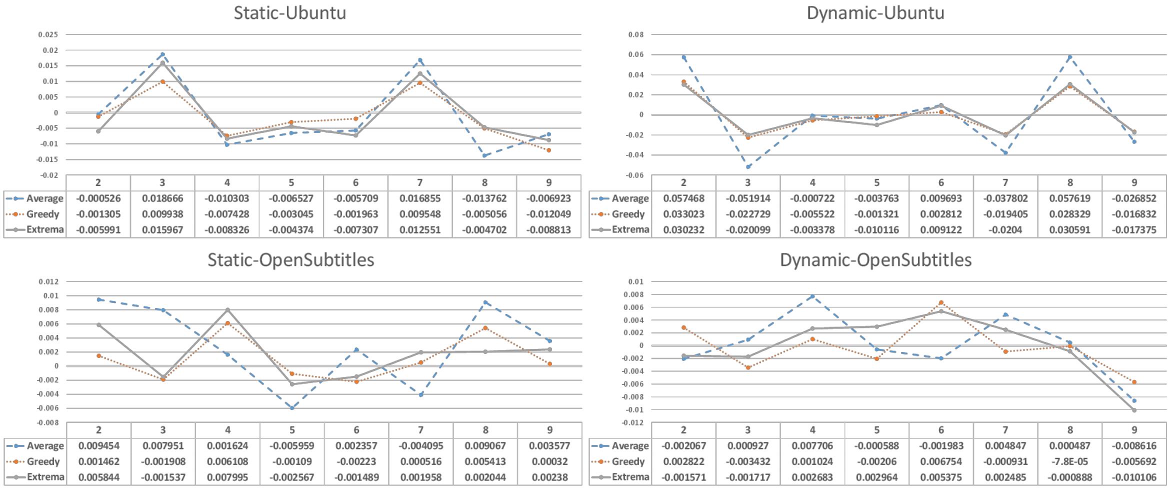
Table 4: The results of human evaluation on Ubuntu and OpenSubtitles datasets.

conversation. **Coherence** is thus used to evaluate the semantic consistency between a generated response and its context. The Coherence score is in the range of 0 to 2, where 0, 1, 2 denote *incoherent*, *neutral* and *coherent*, respectively. In some cases, a coherent response may not be a natural one. Given an example message, "*Can you tell me the way to the nearest bazaar?*", the response "*Yes, I can tell you the way to the nearest bazaar.*" is definitely a coherent but not a natural response. A more extreme example of a message-response pair is "*I don't know what you are talking about!*" and "*I don't know what you are talking about!*". Therefore, besides the Coherence, we proposed another metric, Naturalness, to evaluate the quality of generated responses. For human evaluation, given a context and a conversational response generated by a model, **Naturalness** denotes whether the response can be an alternative to a human response. The Naturalness value equals to 1 or 0, which represents the generated response can be an alternative to a human response or not, respectively. Besides the Coherence and Naturalness, we also want to compare the **Diversity** of the responses generated by all baselines and our proposed approach. Here, diversity score of a generated response equals to the number of distinct tokens in the response divided by the total number of distinct tokens in its context (including the number of distinct tokens in the response). The final Diversity score is the average diversity of all the generated responses in test set.

In the human evaluation, for each model, we randomly sample 500 test results from Ubuntu and OpenSubtitles datasets, respectively. Each of the three annotators, who are undergraduate students and not involved in other parts of the experiment, is asked to provide the evaluation scores for all the 8,000 test results. The final score of each model equals to the average score of the three annotators. Table 4 shows the human evaluation results on the two datasets. Generally speaking, we can see that the proposed static attention-based model outperforms the baselines in Coherence and Naturalness on Ubuntu dataset and obtains comparable performance with the HRAN model in Coherence on OpenSubtitles dataset. For the Diversity, we can see that the proposed dynamic attention-based model is better at generating diverse responses than other models on Ubuntu dataset. We also notice that the CVAE model obtains the best diversity performance on OpenSubtitles dataset and the best Naturalness performance on Ubuntu dataset.

3.2.3 Analysis of Context Length

To verify the impact of context length on the performance of the proposed model for the generation of conversational responses, we use different length of context to re-train the proposed models, which are called context varied models, on two datasets. Here, context length indicates the number of historical utterances that are used for encoding in a context. Figure 3 shows that the performance of the proposed static and dynamic attention models are varying with the change of context length. The values denote the difference between the results of Static$_{\rightarrow}$ and Dynamic$_{\rightarrow}$ in Table 3 and the results of the context varied models. It also verifies that the generation of conversational responses is a context-sensitive process, which relates to the numbers of utterance in context for encoding. Table 5 shows the conversational responses, which are sampled from the test result generated by the proposed static attention model. We can see that the attention values predicted by the static attention model can appropriately reveal the importance of the utterance in a context for generating conversational responses.



Static-Ubuntu

	2	3	4	5	6	7	8	9
Average	-0.000526	0.018666	-0.010303	-0.006527	-0.005709	0.016855	-0.013762	-0.006923
Greedy	-0.001305	0.009938	-0.007428	-0.003045	-0.001963	0.009548	-0.005056	-0.012049
Extrema	-0.005991	0.015967	-0.008326	-0.004374	-0.007307	0.012551	-0.004702	-0.008813

Dynamic-Ubuntu

	2	3	4	5	6	7	8	9
Average	0.057468	-0.051914	-0.000722	-0.003763	0.009693	-0.037802	0.057619	-0.026852
Greedy	0.033023	-0.022729	-0.005522	-0.001321	0.002812	-0.019405	0.028329	-0.016832
Extrema	0.030232	-0.020099	-0.003378	-0.010116	0.009122	-0.0204	0.030591	-0.017375

Static-OpenSubtitles

	2	3	4	5	6	7	8	9
Average	0.009454	0.007951	0.001624	-0.005959	0.002357	-0.004095	0.009067	0.003577
Greedy	0.001462	-0.001908	0.006108	-0.00109	-0.00223	0.000516	0.005413	0.00032
Extrema	0.005844	-0.001537	0.007995	-0.002567	-0.001489	0.001958	0.002044	0.00238

Dynamic-OpenSubtitles

	2	3	4	5	6	7	8	9
Average	-0.002067	0.000927	0.007706	-0.000588	-0.001983	0.004847	0.000487	-0.008616
Greedy	0.002822	-0.003432	0.001024	-0.00206	0.006754	-0.000931	-7.8E-05	-0.005692
Extrema	-0.001571	-0.001717	0.002683	0.002964	0.005375	0.002485	-0.000888	-0.010106

Figure 3: The impact of different number of utterances, used for encoding, in a context (context length) on the performance of the proposed static and dynamic attention based models on two experimental datasets. The range of context length is from 2 to 9.

Sample 1		α
Context	Bonhoeffer, what drives do you want to mount?	0.4872
	What?	0.1535
	I have an ext3 usb drive.	0.3593
Response	I have a usb drive, I want to use it for testing.	

Sample 2		α
Context	Are you using a 32bit version of firefox? why?	0.2284
	Need to use webex, doesn't support 64bit java.	0.4660
	Using 64bit firefox.	0.3056
Response	I have a 64-bit machine, and I want to use 64 bit.	

Table 5: Samples of the generated responses from the Ubuntu test result. α indicates the attention value of each utterance in context calculated by the proposed static attention mechanism.

4 Related Work

Ritter et al. (2010) proposed an unsupervised approach to model dialogue response by clustering the raw utterances. They then presented an end-to-end dialogue response generator by using a phrase-based statistical machine translation model (Ritter et al., 2011). Banchs and Li (2012) introduced a search-based system, named IRIS, to generate dialogues using vector space model and then released the experimental corpus for research and development (Banchs, 2012). Recently, benefit from the advantages of the **sequence-to-sequence learning** framework with neural networks, Sutskever et al. (2014) and Shang et al. (2015) had drawn inspiration from the neural machine translation (Bahdanau et al., 2014) and proposed an RNN encoder-decoder based approach to generate dialogue by considering the last one sentence and a larger range of context respectively. Serban et al. (2016b) proposed a parallel stochastic generation framework which first generates a coarse sequence and then generates an utterance conditioned on the coarse sequence. Shao et al. (2017) introduced the "glimpse-model" which adds self-attention to the decoder to maintain coherence for generating long, informative, coherent and diverse responses in single turn setting. Yao et al. (2017) first predicted cue words using point-wise mutual information (P-MI) for short text conversation generation and then added them into the encoder-decoder framework. To consider the **context information** for improving the diversity of generated conversations, Serban et al. (2016a) presented a hierarchical recurrent encoder-decoder (HRED) approach to encode each utterance and recurrently model the dialogue context to generate context-dependent responses. Serban et al. (2017b) further introduced a stochastic latent variable at each dialogue turn to improve the ambiguity and uncertainty of the HRED model for dialogue generation. Xing et al. (2017) proposed a hierarchical recurrent attention network (HRAN) to jointly model the importance of tokens in utterances and the utterances in context for context-aware response generation. Tian et al. (2017) presented a context-aware hierarchical model to generate conversations by jointly modeling the utterance and inter-utterance information for encoding process. As the advantages of **generative adversarial net (GAN)** and **variational autoencoder (VAE)**, Yu et al. (2017) proposed a sequence generative adversarial net model to assess a partially generated sequence with policy gradient and obtain the intermediate rewards by using Monte Carlo search. Zhao et al. (2017) modified the VAE model by conditioning the response into the VAE model in training step to optimize the similarity of prior network and recognition network for dialogue generation. Similarly, Shen et al. (2017) presented a conditional variational framework to generate specific responses based on the dialog context. Due to the recent advantages of **reinforcement learning** on modeling human-computer interactions, such as the AlphaGo (Silver et al., 2016), researchers begin to focus on modeling the success of a conversation by not only considering the quality of single turn response generation, but also considering long-term goal of the conversation. To address the problems of generating generic and repetitive response of the RNN encoder-decoder framework, Li et al. (2016c) proposed a deep reinforcement learning approach to either generate meaningful and diverse response or increase the length of the generated dialogues. Dhingra et al. (2017) presented an end-to-end dialogue system for information accquisition, which is called KB-InfoBot from knowledge base (KB) by using reinforcement learning. Asghar et al. (2017) proposed an active learning approach to learn user explicit feedback online and combine the offline supervised learning for response generation of conversational agents.

5 Conclusion and Future Work

This paper proposed a novel context-sensitive generation approach for open-domain conversational responses. The proposed model gained from the proposed static and dynamic attention for context or inter-utterance representation. Experimental results show that the proposed model generally outperforms all the baselines in automatic and human evaluations. It is also verified the impact of context length on the performance of the proposed generation models for conversational responses. In future work, the way to uniformly integrate the static and dynamic attention for decoding will be explored.

Acknowledgements

The authors would like to thank all the anonymous reviewers for their insightful comments. The paper is supported by the NSFC (No. 61502120, 61472105, 61772153).

References

Nabiha Asghar, Pascal Poupart, Xin Jiang, and Hang Li. 2017. Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 78–83, Vancouver, Canada, August. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Rafael E. Banchs and Haizhou Li. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *ACL*, pages 37–42.

Rafael E. Banchs. 2012. Movie-dic: a movie dialogue corpus for research and development. In *ACL*, pages 203–207.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the*

55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 484–495. Association for Computational Linguistics.

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

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic, September. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *NIPS*, 26:3111–3119.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *NAACL*, pages 172–180.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *EMNLP*, pages 583–593.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models.

Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. 2016b. Multiresolution recurrent neural networks: An application to dialogue response generation. *arXiv preprint arXiv:1606.00776*.

Iulian Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. 2017a. Multiresolution recurrent neural networks: An application to dialogue response generation.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues. pages 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 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*, pages 2200–2209. Association for Computational Linguistics.

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A conditional variational framework for dialog generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 504–509. Association for Computational Linguistics.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484.

Ilya Sutskever, Oriol Vinyals, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *NIPS*, 4:3104–3112.

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–236. Association for Computational Linguistics.

J Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*.

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

Chen Xing, Wei Wu, Yu Wu, Ming Zhou, Yalou Huang, and Wei-Ying Ma. 2017. Hierarchical recurrent attention network for response generation. *arXiv preprint arXiv:1701.07149*.

Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. 2017. Towards implicit content-introducing for generative short-text conversation systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2180–2189. Association for Computational Linguistics.

L Yu, W Zhang, J Wang, and Y Yu. 2017. Seqgan: sequence generative adversarial nets with policy gradient. In *AAAI Conference on Artificial Intelligence, 4-9 February 2017, San Francisco, California, Usa*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 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)*, pages 654–664. Association for Computational Linguistics.