Knowledge Grounded Pre-Trained Model For Dialogue Response Generation

Yanmeng Wang[1, 2], Wenge Rong[1, 2], Jianfei Zhang[1, 2], Yuanxin Ouyang[1, 2], Zhang Xiong[1, 2]
[1]State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China
[2]School of Computer Science and Engineering, Beihang University, Beijing 100191, China
{wang.ym, w.rong, zhangjf, oyyx, xiongz}@buaa.edu.cn

***Abstract*—Teaching machine to answer arbitrary questions is a long-term goal of natural language processing. In real dialogue corpus, informative words like named entities can often be infrequent and hard to model, and one primary challenge of dialogue system is how to promote the model's capability of generating high-quality responses with those informative words. In order to address this problem, we propose a novel pre-training based encoder-decoder model, which can enhance the multi-turn dialogue response generation by incorporating external textual knowledge. We adopt BERT as encoder to merge external knowledge into dialogue history modeling, and a multi-head attention based decoder is designed to incorporate the semantic information from both knowledge and dialogue hidden representations into decoding process to generate informative and proper dialogue responses. Experiments on two response generation tasks indicate our model to be superior over competitive baselines on both automatic and human evaluations.**

***Index Terms*—Multi-turn Dialogue, Response Generation, Pre-trained Model, Unstructured Knowledge**

I. Introduction

Generating appropriate answers to arbitrary questions in an automated dialogue system is a difficult and challenging task [1]. To obtain meaningful, smooth and grammatically compliant responses, plenty of approaches have been proposed and can be roughly divided into two categories, i.e., retrieval based and generative based models [2]. Currently the latter one has been attached much attention in the community and plays an important role in the multi-turn dialogue systems.

Earlier solutions for response generation models include statistical phrase based machine translation [3]. Later on, the neural network, e.g., Seq2Seq [4] and its variants, has substantially advanced the state-of-the-art in dialogue generation and is able to give fluent and grammatical responses. For example, the hierarchical recurrent encoder-decoder (HRED) has been proposed with latent variable models to generate more informative responses [5], [6].

Most recent generative models are trained on large corpora of multi-turn dialogues. However, it is found that learning semantic interactions merely from dialogue corpus may be not enough [7], since many of the entities are sparsely represented in existing conversational datasets. For example, Ubuntu Dialogue Corpus [5] is a widely used technical oriented conversation corpus and it contains many Ubuntu program names and commands in the multi-turn dialogues. However, the dialogues do not explicitly contain the functioning and usage of Ubuntu entities involved, which is critical information for dialogue model to produce informative and coherent responses. Some researchers have argued to inject external information to overcome this problem [7], [8]. For example, Wang et al. employed unstructured external knowledge to facilitate more informative response generation [9].

At the same time, the pre-trained language models [10], [11] have recently gained significant performance improvement in many natural language understanding tasks. Due to the fully connected self-attention deep structure and large amounts of unlabeled data, the pre-trained models are able to capture complex semantic information of long-form language, which is important for multi-turn dialogue understanding and response generation. In parallel to dialogue generation with unstructured knowledge enhancement, another closely related line of research is document grounded conversations generation, which aims to generate dialogue responses when chatting about the content of a specific document [12]. This task also requires to integrate document knowledge with multi-turn dialogue history.

Therefore, in this paper we propose an advanced dialogue response generation framework by facilitating the external knowledge fusion and the capability of pre-trained language model. The dialogue history and text description of technical entities involved in dialogue are fed into a novel Bert2trans encoder-decoder model to incorporate the semantic information from both text knowledge and dialogue history. In this work, we use BERT [10] as the encoder which is able to fuse the information by its multiple layers of self-attention in both right-to-left and left-to-right direction. The BERT based encoder can provide better dialogue semantic representation and history understanding. The pre-trained multiple layers of self-attention fuse the information from dialogue history and augment the dialogue semantic representation with the given unstructured knowledge for better dialogue history understanding. Furthermore, we also adopted transformer decoder (multi-head attention) to leverage the fused dialogue and knowledge representations in sequential or joint way to generate proper and informative response to the multi-turn dialogue.

The framework is evaluated on both Ubuntu dialogue [1] and Document Grounded Conversations [12] task. Experiment results show that our model is capable to retrieve relevant information from the description of matched Ubuntu entities

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

and the background document to generate informative and coherent responses. Both automatic and human evaluations show that our model substantially outperforms the competitive baselines.

The key contributions of our work are three-folds: 1) we propose a novel Bert2trans encoder-decoder model leveraging pre-trained language models to encode the semantic information from both unstructured text and conversation history, which is applicable to both Ubuntu response generation with unstructured knowledge enhancement and document grounded conversation generation task. 2) we adopted transformer decoder (multi-head attention) to leverage the fused dialogue and knowledge representations in sequential or joint way to generate proper and informative response to the multi-turn dialogue. 3) we evaluate our framework on two datasets. Our model significantly outperforms competitive baselines and achieves state-of-the-art on two generation tasks, Ubuntu Dialogue Response Generation task and Document Grounded Conversations task.

II. Related Work

Automatic dialogue response generation has attracted considerable interest in the community. Recently Seq2Seq model has been widely adopted in the field of dialogue response generation [13], [14]. Gu et al. argued that it would be beneficial that Seq2Seq system can accommodate both understanding and copy mechanism in case that system needs to refer to some words of target side sentence [15]. It is difficult for Seq2Seq system to learn the meaning of the rare words such as proper nouns and to generate them with standard RNN model. Therefore, several researchers extended attention-based encoder-decoder with CopyNet (Pointer network), which predicts words based on combined probability distribution [15]–[17]. The probability distribution for the vocabulary can be obtained from softmax layer of conventional Seq2Seq model, while the probability distribution for the input sequence can be obtained by Pointer Networks. Therefore, it is possible for the model to copy some words directly from the input text into the output. It is important for conversational model in response generation, as the response frequently repeats sub-sequence of the context [18].

To overcome the long propagation problem in model training, hierarchical recurrent encoder-decoder (HRED) [5] was proposed with two-level encoders for better dialogue context understanding. The word-level RNN encodes all tokens in each dialogue turn into utterance vectors. The context-level encoder recursively summarizes the dialogue turns into hidden states as representation of previous dialogue context, which is fed into decoder as condition to predict next turn utterance. Zhao et al. presented an encoder-decoder framework based on conditional variational autoencoders (VAE) that captures the discourse-level diversity in the encoder [6]. The VAE is able to encode the contextual utterances into a probabilistic distribution instead of a point encoding. This allows the model to generate diverse responses by drawing samples from the learned distribution and reconstruct their words via a decoder neural network.

Beside improving the Seq2Seq model itself, another line of research is to incorporate external knowledge into dialogue model. For example, Li et al. represent user personality in a learnable embedding and generate personality coherent responses [19]. The people background information (such as address and nationality) works as extra labels in the training process. Kottur et al. further propose a neural generative dialog model conditioned on both speakers and context history to promote response diversity [20]. Considering that relevant facts can significantly affect response generation, Ghazvininejad et al. present a Seq2Seq model which can incorporate external relevant knowledge. Beside the dialogue context encoder, a variance of memory network was adopted to encode multiple textural facts into vector representations [8].

In addition, large-scale unsupervised pre-trained language model, e.g., BERT [10], XLNet [11], RoBERTa [21] have brought significant performance gains over previous RNN methods in many NLP tasks, such as reading comprehension, sentence classification, and question answering. Some recent literature also proposed to leveraging pre-trained language models in text generation task [22]. In this work, we use BERT on the encoder which is able to fuse the information by its multiple layers of self-attention in both right-to-left and left-to-right direction. Furthermore, we also adopt transformer decoder (multi-head attention) to leverage the fused dialogue and knowledge representations to generate appropriate and informative response to the multi-turn dialogue.

This work distinguishes itself from previous work which enhance RNN-based Seq2Seq framework with memory network, whereas we are attempting to leverage pre-trained language model to fuse dialogue history with relevant external knowledge by multi-level attention mechanism on dialogue session level. We adopt transformer decoder (multi-head attention) with two different decoder mechanisms to fuse the dialogue and knowledge representations in sequential or joint way to generate response to the multi-turn dialogue.

III. Methodology

A. Overview

The proposed knowledge enhanced dialogue framework is presented in Fig. 1, in which a novel encoder-decoder model adopts BERT with knowledge indication embedding as encoder to fuse the external text knowledge with dialogue history. A multi-head attention based decoder is proposed to incorporate the semantic information from both encoded knowledge and dialogue hidden states into decoding to generate informative and coherent dialogue responses.

We use $D = \{x_1, ..., x_m\}$, $R = \{y_1, ..., y_n\}$, and $K = \{s_1, ..., s_l\}$ to represent dialogue history, response and relevant text knowledge respectively, x_t, y_t, and s_t are the words from vocabulary V. For a given dialogue history D and corresponding text knowledge K, our goal is to generate a proper and informative next turn utterance R.

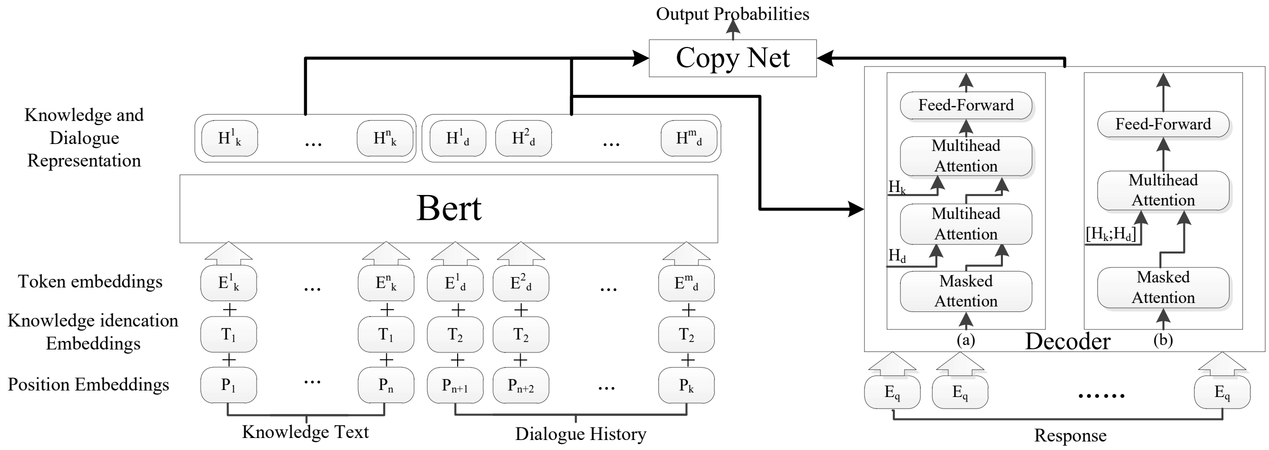

Fig. 1. Knowledge Grounded Response Generation

B. Encoder

As most response generation models have an encoder-decoder structure [2], [13], we adopt BERT model [10] to map the input sequence of dialogue history and relevant knowledge into a sequence of continuous representation. The words in dialogue history D are used as key to match entities with related document. We pack the descriptions of matched entities into a text sequence as K. The dialogue history is concatenated with knowledge text and $[SEP]$ is inserted between, i.e., $\{[CLS]\ K\ [SEP]\ D\ [SEP]\}$.

Apart from position and token embedding, we also add a learned knowledge embedding to every token indicating whether it belongs to the knowledge text or dialogue history. As shown in Fig. 1, for each token x_i, the input embedding is the sum of its token embedding, knowledge indicating embedding and position embedding:

$$I(x_i) = E(x_i) + T(x_i) + P(x_i) \tag{1}$$

where $E(x_i)$, $T(x_i)$ and $P(x_i)$ are word embedding, knowledge indication embedding and position embedding, respectively. The input embeddings are then fed into BERT model to get the knowledge and dialogue history encoding representations.

$$H_k; H_d = BERT(I(k_1)...I(k_l), I(x_1)...I(x_m)) \tag{2}$$

C. Decoder

After comprehensive fusion in BERT model, H_k and H_d are the semantic representations of knowledge text and dialogue history respectively. As illustrated in Fig. 1, we propose two different decoder mechanisms.

$$D_m = MultiHead(Q = R, K = R, V = R) \tag{3}$$

where R is the response embedding and D_m is the output of masked self-attention. Following definition proposed in [23], Q, K, V denote the query, key, and value vectors in multi-head attention, respectively.

We propose two decoders to fuse H_k and H_d with different strategies. Firstly, as shown in part (a) in Fig. 1, we apply two separate multi-head attentions to encode the hidden states of dialogue history and knowledge text sequentially:

$$D_d = MultiHead(Q = D_m, K = H_d, V = H_d) \tag{4}$$

$$D_{kd} = MultiHead(Q = D_d, K = H_k, V = H_k) \tag{5}$$

where H_d and H_k are fused semantic representations of dialogue history and knowledge text, which are the direct output from BERT encoder. D_d and D_{kd} are the response representations after fused with dialogue history H_d and knowledge text H_k.

Meanwhile, in another decoder (illustrated in part (b) in Fig. 1), the H_d is concatenated with H_k and fused with response representation in multi-head attention:

$$D_{kd} = MultiHead(Q = D_m, K = [H_k; H_d], V = [H_k; H_d]) \tag{6}$$

Comparing the two decoders, decoder (a) can better recognize the border of knowledge text and dialogue history, while (b) jointly fuses the information with complete vision from both H_k and H_d. Next, the D_{kd} was fed into Position-wised Feed-Forward Networks.

$$F_{dk} = FFN(D_{dk}) \tag{7}$$

D. Copy Mechanism

We apply dot-product attention [24] to hidden states F_{dk} output from decoder with encoding representations $[H_k; H_d]$ to obtain attended token representations $F^{'}_{dk}$ and corresponding attention weight distribution W, and the vocabulary distribution M is calculated subsequently via a feedforward neural network:

$$F^{'}_{dk}, W = Attention(F_{dk}, [H_k; H_d]) \tag{8}$$

$$M = softmax(FNN(F^{'}_{dk})) \tag{9}$$

Copy mechanism [15], [25] is introduced to allow both copying words from input sequence and generating words from a pre-defined vocabulary V during decoding. In this paper, we concatenate the attended token representations $F^{'}_{dk}$ and the outputs of decoder F_{dk} to learn the generation probability P_{gen}, which is used to obtain the final distribution of generated response R:

$$P_{gen} = sigmoid(FNN(F_{dk}^{'}; F_{dk})) \tag{10}$$

$$R = P_{gen}M + (1 - P_{gen})W \tag{11}$$

where M is the vocabulary distribution from Equation 9 and W is the attention weight distribution from Equation 8.

IV. Experimental Study

A. Datasets

We evaluate our proposed knowledge enhanced response generation model on Ubuntu Dialogue Corpus [1] and Document Grounded Conversations Dataset [12].

1. Ubuntu Dialogue Corpus. Most Ubuntu dialogues are goal-oriented, i.e., one user posts a specific technical question and others try to help solve the problem. The nature of the Ubuntu corpus is suitable for exploring models for generating informative responses. In order to evaluate our proposed model, we remove the dialogues which do not contain any Ubuntu entity. 272,128 dialogues in training corpus are retained, as well as 11,594 dialogues in test set and 11,880 dialogues in validate set. On average, there are 5.5 turns per dialogue and the detailed statistics of this dataset is given in Table I.

TABLE I
Statistics of Datasets.

	Ubuntu		Doc Ground	
	Train	**Test**	**Train**	**Test**
Dialogs	272k	12k	73k	12k
Turns	1.5M	62k	264k	40k
Tokens	32.6M	1.5M	5.2M	0.7M
Entity	0.75M	33.6k		

We crawl Ubuntu manual pages to build the knowledge database, which contains 76k Ubuntu-related entities, including commands and system programs. The descriptions of Ubuntu-related entity are retrieved as the external knowledge. By text matching, we initially select 5,218 entities which exits in dialogue corpus. However, many of matched entities are commonly used English words (e.g., “the” or “I”). To avoid mismatching, we remove all English words from matched entities and finally 3,715 entities are retained.

2. Document Grounded Conversation. We used processed version from [12], which contains 72922 training dialog samples and 11577 test samples. A related document is given for each dialogue, which may contain descriptive information about the movie.

B. Evaluation Metrics

On the Ubuntu dataset, we employ three evaluation metrics to measure the response generation accuracy [1], i.e., BLEU [26], METEOR [27] and ROUGE-L [28], which have been widely used for measuring textual similarity. Following previous studies [1], [4], we also perform evaluations using the distinct-1, distinct-2 [4] to measure the abilities of our proposed models to promote diversity in dialogue generation. We further use average information entropy and entity number per response [1] to measure the performance that can provide more informative content in response generation. On Document Grounded Conversation dataset, we follow [12] and adopt perplexity (PPL) and BLEU [26] for comparison.

Moreover, in this paper, we adopt human evaluation, since it is broadly agreed that objective metrics weakly correlate with human evaluation results. Human evaluation is a necessity in dialogue generation.

C. Training Details

TABLE II
Hyperparameters for Model Training

Hyperparameters		Ubuntu	DGC
Encoder	Number of Layers	12	12
	Hidden size	768	768
	Attention heads	12	12
	FFN inner hidden size	3072	3072
Decoder	Number of Layers	2	2
	Hidden size	768	768
	Attention heads	8	8
	FFN inner hidden size	2048	2048
Passage max length		460	260
Conversation max length		49	120
Target max length		30	60
Training batch size		48	30
Learning rate		5e-5	5e-5
Beam search size		5	5
Vocab size		30522	30522

Table II presents the hyperparameters for training our proposed model. We adopt pre-trained BERT$_{base}$ in our proposed models. The Bert2trans decoder has 2 layers of attention blocks. All models are trained using Adam [29] for optimization. The learning rate is set to 1e-4 for RNN-based baseline model and 5e-5 for our proposed model. All the models are trained at most 20 epochs.

D. Experimental Results

On Ubuntu dataset, we make three groups of experiment settings:

1) RNN-based baselines: The LSTMLM baseline follows the settings described in [5] with 500 hidden units for single LSTM layer. The HRED baseline follows the same settings as described in [5] with 500, 1000 and 500 hidden units respectively for the encoder RNN, the context RNN and decoder RNN. The mini-batch size is set to 80. The encoder RNN is a standard GRU structure. VHRED extends the HRED model by introducing random variables in context RNN.

2) Attention-base baselines: Transformer [23] and Transformer enhanced with copynet [15].

3) Our proposed models: Bert2trans+kg stands for BERT to transformer decoder enhanced by external textual knowledge. As described in subsection III-C, we adopt two different strategies, which are denoted as Bert2trans+kg$_a$ and Bert2trans+kg$_b$.

Table III presents results of multiple evaluation metrics for our models and the baselines. Compared with other baselines,

TABLE III
RESULTS ON UBUNTU TEST SET

	BLEU	**METEOR**	**ROUGE-L**
LSTMLM	0.0745	0.0135	4.17
VHRED	0.4114	0.0272	8.08
HRED	0.7756	0.0340	10.53
Transformer	1.1982	0.0364	8.12
Transformer+copy	1.2724	0.0378	8.78
Bert2trans+kg$_a$	1.6879	0.0388	13.03
Bert2trans+kg$_b$	**1.7633**	**0.0425**	**13.13**

our proposed models achieve much improvement on all the three metrics. The pre-trained Bert2trans+kg model is capable of generating more contextualized tokens. Incorporating external knowledge can enhance the model capacity of capturing informative features, leading to longer and more meaningful responses. Besides, we could observe that the Bert2trans+kg$_b$ model slightly outperforms Bert2trans+kg$_a$ model.

On Document Grounded Conversation dataset, we make three groups of experiment settings:

1) None-Knowledge baselines: The Seq2Seq with attention [24]. The HRED baseline follows the same settings as described in [5] with 500, 1000 and 500 hidden units respectively for the encoder RNN, the context RNN and decoder RNN. Transformer: The competitive NMT model based on multi-head attention [23]. Models in this group only take dialogue history as input and don't use document knowledge.

2) Knowledge-grounded baselines: Seq2Seq + knowledge and HRED + knowledge extend the Seq2Seq and HRED framework to leverage the corresponding document knowledge respectively. They use same encoder to encode both the dialog and document knowledge. Please refer to [30] for more details. Wizard Transformer is a Transformer-based model for multi-turn dialogue with dialogue history and unstructured text knowledge concatenated as input. The ITE+CKAD [12] and ITE+DD [12] stand for Incremental Transformer Encoder (ITE) as encoder and Context Knowledge-Attention Decoder (CKAD) and Deliberation Decoder (DD) respectively, which achieved state-of-the-art results on Document Grounded Conversation task. Please refer to [12] for more details.

3) Our proposed models: Bert2trans+kg$_b$ stands for BERT to transformer decoder enhanced by external textual knowledge with concatenated hidden representation for dialogue history and text knowledge. In order to make an aligned comparison with [12], we also equip our proposed Bert2trans+kg$_b$ model with Deliberation Decoder as [12].

Table IV illustrates PPL and BLEU results on Document Grounded Conversation dataset. The results of baselines are reported from [12]. The ITE+DD model performs best among all the baselines, which adopts deliberation decoder [12], [22] to enhance the model. Results have shown that our proposed Bert2trans+kg$_b$ significantly outperforms all baseline model on BLEU scores. The combination of Deliberation Decoder only remarkably improves the performance on PPL. Our Bert2trans+kg+DD model outperforms all baseline models on both BLEU and PPL and achieved a new state-of-the-art result on Document Ground Conversation Task [12].

TABLE IV
PPL AND BLEU ON DOCUMENT GROUNDED CONVERSATION

	PPL	**BLEU**
Seq2Seq	80.93	0.38
HRED	80.84	0.43
Transformer	87.32	0.36
Seq2Seq+kg	78.47	0.39
HRED+kg	79.12	0.77
Wizard Transformer	70.30	0.66
ITE+CKAD	64.97	0.86
ITE+DD	15.11	0.95
Bert2trans+kg$_b$ (ours)	39.89	1.14
Bert2trans+kg$_b$+DD (ours)	**10.52**	**1.18**

TABLE V
GENERATION DIVERSITY RESULTS ON UBUNTU TEST SET

	dct-1	dct-2	$\|1gram\|$	$\|wd\|$
LSTMLM	0.022	0.080	1138	53k
VHRED	0.020	0.093	2176	107k
HRED	0.012	0.051	1529	128k
Transformer	0.010	0.043	1596	168k
Transformer+copy	0.017	0.080	2806	162k
Bert2trans+kg$_a$	**0.024**	**0.108**	4104	172k
Bert2trans+kg$_b$	**0.024**	0.104	**4368**	**180k**

Existing Seq2Seq models tend to generate generic non-informative response. We present the comparison of response diversity results on Ubuntu test set in Table V. The dct-1 and dct-2 refer to distinct-1 and dictinct-2 respectively proposed by [4], which are computed as the number of distinct unigrams and bigrams divided by total number of generated words. In table V, Distinct unigrams and total number of generated words are also illustrated as $|1gram|$ and $|wd|$. Our proposed Bert2trans+kg model substantially increases dct-1, dct-2 and the distinct unigram number over transformer+copy models, which indicates that our proposed models can significantly improve the diversity of generated response over competitive baselines.

TABLE VI
RESPONSE INFORMATION ENTROPY AND ENTITY NUMBER ON UBUNTU DATASET

	$\|U\|$	H$_w$	**H**$_U$	Ent
LSTMLM	4.53	6.01	27.24	0.05
VHRED	9.23	6.71	62.01	0.21
HRED	11.08	6.24	69.12	0.22
Transformer	14.51	5.58	80.98	0.31
Transformer+copy	13.96	5.99	83.56	0.32
$Bert2trans + kg_a$	14.87	6.39	95.00	**0.37**
$Bert2trans + kg_b$	**15.52**	**6.41**	**99.46**	0.36

Following experiment in [1], we present the average response length, entity numbers and average response information entropy with respect to the maximum likelihood unigram model over the generated responses in Table VI. The $|U|$ refers to average response length. The H_w refers to

TABLE VII
EXAMPLES OF GENERATED DIALOG RESPONSES ON UBUNTU DATASETS

Case 1:
Menu page: **lvm:** lvm provides the command-line tools for lvm2 **fstab:** the file fstab contains descriptive information about the various file systems **fdisk:** is a disk partition manipulation program, allows you to create, destroy, resize, move and copy partitions on a hard drive using a menu-driven interface **uuid:** the uuid library is used to generate unique identifiers for objects that may be accessible beyond the local system **lshw:** lshw is a small tool to extract detailed information on the hardware configuration of the machine
Dialog history: T1: never tried mounting lvm on a separate machine . I'm sure there 's a way , just haven't tried it .specify the correct file type using -t parameter T2: in the fstab file ? T3: mount -t . add your special ingredients to that command . in fact , " man mount " . try it first using " mount ", then modify your fstab according to what worked with mount. you may be specifying the wrong fs type , or you specified the entire drive when in fact you want to mount a partition T4: yes thats te error message i am receiving inspect the partitions on that drive using fdisk -l . T5: my guess is that your drive is partitioned , and you're trying to mount /dev/sda instead of /dev/sda1 (as an example) if i do fdisk -l nothing happends (do i need to type the hdd behind it ? T6: yes , that 's if you reference the correct uuid. i see the drive with fdisk -l (/dev/sdc1 system GPT) but i dont see the filesystem it is. **Responses:** **Gold:** i see the drive with fdisk -l (/deve/sdc1 system gpt), but I dont see the filesystem it is **Bert2Trans+kg$_b$:** so if I do fdisk -l /dev/sdb, i can see the uuid ? **Transformer+copy:** ok, i'll try that. thanks. **HRED:** how do I do that ?
Case 2:
Menu page: **install**: install program copies files (often jest compliled) into destination locations you choose **xorg:** xorg is a full featured x server that was originally designed for unix and unix - like operating systems running on intel x86 hardware **chroot:** run command with root directory set to newroot **lspci:** lspci is a utility for displaying information about pci buses in the system and devices connected to them **lshw:** lshw is a small tool to extract detailed information on the hardware configuration of the machine
Dialog history: T1: The system may be poweroff or perhabs external connecton is refussed or the system has no openssh-client installed at that system if it is debian based Or may be due to iptable rules , Pardon me if i am wrong T2: the machine is right next to me and can ssh my machine . T3: it means that machine has set iptables for no ssh connection also there might be a possibility that , that machine have no openssh-server installed that is necessary T4: what should I do then ? T5: Go and install openssh-server if already not nistalled in that system and then allow your system ip for ssh to that machine using iptables , you can flush iptables but it is not good idea , use a rule for ssh. You can also set iptbales rules by ip and also by MAC address T6: I don't understand the "allow your system ip for ssh to that machine using iptables"
Responses: **Gold:** To connect to any machine that machine have must installed openssh-sever and to make a cnnection to that machine you must have to install open-ssh client **Bert2Trans+kg$_b$:** no, openssh-server is not installed by default, you can set it to do ssh-server on mahine 2 and then flush iptables **Transformer+copy:** I don't think so **HRED:** yes

information entropy per word and is computed as $H_w = -\sum_{w \in U} p(w) logp(w)$ [1]. The unigram probabilities are computed according to the maximum-likelihood unigram distribution of the training corpus. The H_U refers to information entropy per response. It can be observed that our models, especially Bert2trans+kg_b, produce response with longer average length and can enhance the utterance entropy H_U, which indicates the capability to generate responses with higher informativeness, which is consistent to the previous experiment about BLEU, METEOR and ROUGE-L scores.

Following [7], we also calculate the number of entities per response to measure the model's ability to generate informative contents related to ubuntu manual page. Many ubuntu entity names (such as "which" and "pip") are also commonly used English words. In order to avoid ambiguity, we made a compact entity list with only non-English-word Entities and count the number of entities per response from the compact list. As shown in Table VI, Bert2trans+kg model generates responses with much more ubuntu entities, which appears to suggest that our proposed knowledge enhanced model is able to promote information-richness in dialogue generation.

E. Case Study

As illustrated in Table VII and VIII, We present some examples from Ubuntu and Document Grounded Conversation dataset. Table VII presents responses generated by our proposed Bert2Trans+kg, Transformer+copy, and HRED models. Transformer and HRED model are not capable of retrieving information from relevant description of Ubuntu entities, while the responses generated by our Bert2trans+kg model are much more informative and relevant to the unstructured knowledge (i.e., description of matched Ubuntu entities in dialogue history). Table VIII show the generated responses on Document Grounded Conversation Dataset by our proposed

TABLE VIII
EXAMPLES OF GENERATED DIALOG RESPONSES ON DOCUMENT GROUNDED CONVERSATION

Case 1:
moive passage: The shape of water, director guillermo del toro, drama cast sally hawkins as elisa esposito, a mute cleaner who works at a scret government laboratory. rating rotten tomatoes: 92% and average: 8.4. The Shape of Water is a 2017 American romantic dark fantasy film directed by Guillermo del Toro and written by del Toro and Vanessa Taylor. It stars Sally Hawkins, Michael Shannon, Richard Jenkins, Doug Jones, Michael Stuhlbarg, and Octavia Spencer. Set in Baltimore, Maryland, in 1962, the story follows a mute cleaner at a high-security government laboratory who falls in love with a captured humanoid creature
Dialog history: T1: Does is ? I haven't read into it much. T2: It got an average of 8.4 in rotten tomatoes T3: That's not bad. What's it about again ? **Responses:** **Gold:** It is about a mute cleaner who was working in governmental laboratory and fell in love with a humanoid amphibian creature. **Bert2Trans+kg$_b$:** It's about a mute custodian who falls in love with a humanoid amphibian creature. **ITDD+DD:** It's fantasy drama
Case 2:
moive passage: Despicable Me is a 2010 American 3D computer-animated action-comedy film produced by Illumination Entertainment as its debut film and project for Universal Pictures. The film was animated by the French animation studio Mac Guff, which was later acquired by Illumination. Directed by Pierre Coffin and Chris Renaud in their directorial debuts from a screenplay by Cinco Paul and Ken Daurio and a story by Sergio Pablos, the title references the main character as he refers to himself and is accompanied by a song by Pharrell Williams.
Dialog history: T1: I like all of the voice actors T2: yeah they were pretty good. did you think it was funny ? T3: it was very funny .. the minions were very cute. do you know what year it came out ? **Responses:** **Gold:** It came out in 2010 **Bert2Trans+kg$_b$:** It came out in 2010 **ITDD+DD:** i thought it was really cute movie

Bert2Trans+kg$_b$ and ITE+DD [12] which achieves best performance on automatic evaluation among baseline models. Both Bert2Trans+kg$_b$ and ITE+DD can output meaningful responses depending on the relevant movie descriptions, however, our Bert2Trans+kg$_b$ model can output more proper response in the context of conversation. For case 1, ITE+DD generates a generic response, while our Bert2Trans+kg$_b$ model correctly predicts the answer text span in the passage and produces appropriate response with more detailed knowledge.

F. Human Evaluation

Automatic evaluation of generative dialog model remains an open research challenge [31]. We also conducted a human evaluation on both Ubuntu and Document Grounded Conversation dataset, as a supplement to the automated metrics. We anonymize the model identities for each generated response. Following [7], we adopt two metrics: appropriateness at content level (whether the response is proper in content, grammar and logic) and informativeness at knowledge level (whether the response offers new knowledge and information beyond the dialog context). We ask human annotators to give a preference in pair-wise comparison, in terms of the two metrics. Tie (e.g., neither of the responses are good) is allowed. The results after removing "Tie" pairs are shown in Table IX and X. Agreements among the annotators are calculated using Fleiss kappa [32].

TABLE IX
HUMAN EVALUATION ON UBUNTU DATASET

	Bert2trans+kg$_b$	vs Trans+copy	Kappa
appropriate	0.652	0.348	0.38
Informative	0.698	0.302	0.39

Three human annotators with Ubuntu system experiments evaluate 100 randomly chosen responses from Bert2trans+kg$_b$ model and make comparison with Transformer+copy. Our proposed Bert2trans+kg model is especially superior over baseline model on informativeness metric (sign test, p-value $<$ 0.005), which suggests that our proposed knowledge enhanced model can better merge external knowledge into dialogue history understanding and is able to generate informative and appropriate responses.

TABLE X
HUMAN EVALUATION ON DOCUMENT GROUNDED CONVERSATION DATASET

	Bert2trans+kg$_b$	vs ITE+DD	Kappa
appropriate	0.661	0.339	0.43
Informative	0.617	0.383	0.41

Three human annotators with a high level of proficiency in English evaluate 100 randomly chosen responses from Bert2trans+kg$_b$ model and make comparison with ITE+DD [12]. As illustrated in Table X, We could observe that our Bert2trans+kg$_b$ model outperforms ITE+DD baseline in terms of both metrics (sign test, p-value $<$ 0.005). Both our proposed model and ITE+DD are capable of retrieving information from the given document and our propose model is able to output more appropriate response in content and logic, which also can be observed through the use cases in Section IV-E. The relatively high kappa scores on Document Grounded Conversation Dataset indicate that annotators reached agreement in most cases.

V. Conclusion

In this research, we proposed a deep pre-trained dialogue generation model, which is enhanced with external textual knowledge to facilitate dialogue understanding and meaningful response generation. We evaluate the model with Ubuntu and Document Grounded Conversation dataset and report the results on multiple metrics, which demonstrate that our approach can generate more coherent and informative dialogue responses.

Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (No. 61977002) and the State Key Laboratory of Software Development Environment of China (No. SKLSDE-2019ZX-16).

References

[1] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. C. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017, pp. 3295–3301.

[2] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, 2015, pp. 1577–1586.

[3] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 583–593.

[4] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "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*, 2016, pp. 110–119.

[5] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016, pp. 3776–3784.

[6] T. Zhao, R. Zhao, and M. Eskénazi, "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*, 2017, pp. 654–664.

[7] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Commonsense knowledge aware conversation generation with graph attention," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 4623–4629.

[8] M. Ghazvininejad, C. Brockett, M. Chang, B. Dolan, J. Gao, W. Yih, and M. Galley, "A knowledge-grounded neural conversation model," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018, pp. 5110–5117.

[9] Y. Wang, W. Rong, Y. Ouyang, and Z. Xiong, "Augmenting dialogue response generation with unstructured textual knowledge," *IEEE Access*, vol. 7, pp. 34 954–34 963, 2019.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[11] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Proceedings of 2019 Annual Conference on Neural Information Processing Systems*, 2019, pp. 5754–5764.

[12] Z. Li, C. Niu, F. Meng, Y. Feng, Q. Li, and J. Zhou, "Incremental transformer with deliberation decoder for document grounded conversations," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019, pp. 12–21.

[13] O. Vinyals and Q. V. Le, "A neural conversational model," *CoRR*, vol. abs/1506.05869, 2015.

[14] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, and B. Dolan, "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*, 2015, pp. 196–205.

[15] J. Gu, Z. Lu, H. Li, and V. O. K. Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1631–1640.

[16] S. He, C. Liu, K. Liu, and J. Zhao, "Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 199–208.

[17] Y.-C. Tam, J. Ding, C. Niu, and J. Zhou, "Cluster-based beam search for pointer-generator chatbot grounded by knowledge," in *Proceedings of 7th Dialog System Technology Challenge*, 2019.

[18] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational AI," in *Proceedings of 56th ACL Tutorial Abstracts*, 2018, pp. 2–7.

[19] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and W. B. Dolan, "A persona-based neural conversation model," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 994–1003.

[20] S. Kottur, X. Wang, and V. Carvalho, "Exploring personalized neural conversational models," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 3728–3734.

[21] Y. Liu, M. Ott, N. Goyal, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[22] H. Zhang, J. Cai, J. Xu, and J. Wang, "Pretraining-based natural language generation for text summarization," in *Proceedings of the 23rd Conference on Computational Natural Language Learning*, 2019, pp. 789–797.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of 2017 Annual Conference on Neural Information Processing Systems*, 2017, pp. 5998–6008.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2015.

[25] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 1073–1083.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2001, pp. 311–318.

[27] K. Mazidi and R. D. Nielsen, "Linguistic considerations in automatic question generation," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 321–326.

[28] L. Chin-Yew, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of 2014 Workshop on Text Summarization Branches Out*, 2014.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[30] K. Zhou, S. Prabhumoye, and A. W. Black, "A dataset for document grounded conversations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 708–713.

[31] Z. Xu, N. Jiang, B. Liu, W. Rong, B. Wu, B. Wang, Z. Wang, and X. Wang, "LSDSCC: a large scale domain-specific conversational corpus for response generation with diversity oriented evaluation metrics," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 2070–2080.

[32] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and Psychological Measurement*, vol. 33, pp. 613–619, 1973.