10-2021

Towards enriching responses with crowd-sourced knowledge for task-oriented dialogue

Yingxu HE
National University of Singapore

Lizi LIAO
Singapore Management University, lzliao@smu.edu.sg

Zheng ZHANG
Sea-NExT Joint Lab

Tat-Seng CHUA
National University of Singapore

Towards Enriching Responses with Crowd-sourced Knowledge for Task-oriented Dialogue

Yingxu He[1], Lizi Liao[1,2], Zheng Zhang[3], Tat-Seng Chua[1]
[1]School of Computing, National University of Singapore
[2]Sea-NExT Joint Lab, Singapore
[3]Department of Computer Science and Technology, Tsinghua University
e0139128@u.nus.edu,liaolizi.llz@gmail.com,zhangz.goal@gmail.com,dcscts@nus.edu.sg

ABSTRACT

Task-oriented dialogue agents are built to assist users in completing various tasks. Generating appropriate responses for satisfactory task completion is the ultimate goal. Hence, as a convenient and straightforward way, metrics such as success rate, inform rate etc., have been widely leveraged to evaluate the generated responses. However, beyond task completion, there are several other factors that largely affect user satisfaction, which remain under-explored. In this work, we focus on analyzing different agent behavior patterns that lead to higher user satisfaction scores. Based on the findings, we design a neural response generation model **EnRG**. It naturally combines the power of pre-trained GPT-2 in response semantic modeling and the merit of dual attention in making use of the external crowd-sourced knowledge. Equipped with two gates via explicit dialogue act modeling, it effectively controls the usage of external knowledge sources in the form of both text and image. We conduct extensive experiments. Both automatic and human evaluation results demonstrate that, beyond comparable task completion, our proposed method manages to generate responses gaining higher user satisfaction.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Discourse, dialogue and pragmatics**.

KEYWORDS

crowd-sourced knowledge, response generation, task-oriented

ACM Reference Format:
Yingxu He, Lizi Liao, Zheng Zhang, Tat-Seng Chua. 2021. Towards Enriching Responses with Crowd-sourced Knowledge for Task-oriented Dialogue. In *Proceedings of the 2nd ACM Multimedia Workshop on Multimodal Conversational AI (MuCAI '21), October 24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3475959.3485392

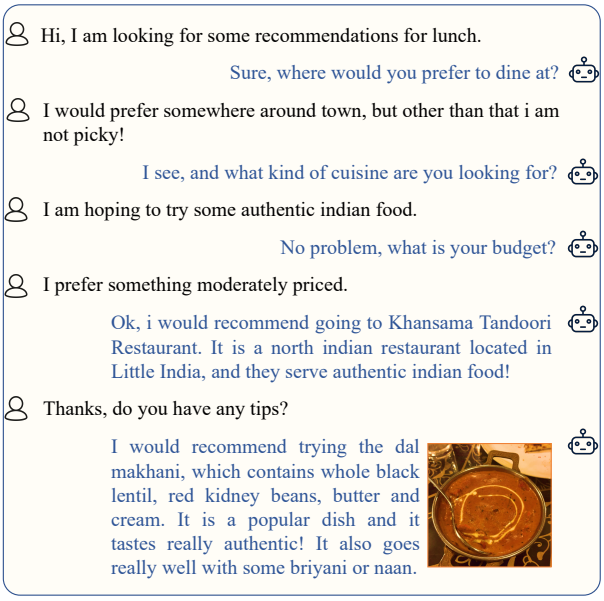


Figure 1: Part of a collected dialogue that contains enriched responses for higher user satisfaction.

1 INTRODUCTION

Task-oriented dialogue systems aim to fulfill the requirements raised by users, such as booking flight tickets or searching for the restaurant through interactions. Due to the huge industry potential, it has attracted a significant amount of attention recently. The development of such agents is often decomposed into three sub-tasks: understanding user inputs, deciding next actions and generating responses. As fulfilling user requirements and enhancing user satisfaction is essential for dialogue agents, the closely related action decision and response generation are usually modeled together and play an important role in the system [8, 33].

Recently, there have been much researches on the response generation sub-task and it is commonly agreed that a good response should fulfill all the requirements from users [1, 34]. Inform rate and success rate are thus the popular metrics to reflect the ability of the dialogue systems of handling queries. Metrics such as BLEU score [24] and ROUGE [17] are also widely adopted in both open-domain and task-oriented dialogues. In general, the former ones evaluate the agent's performance on task completion, while the latter ones judge the literal overlap between generated responses and ground truth responses.

However, although the above-mentioned metrics are successful in evaluating the completeness of tasks and the surface realization of text, it lacks in reflecting the intelligence of dialogue systems and

MuCAI '21, October 24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8679-1/21/10…$15.00
https://doi.org/10.1145/3475959.3485392

satisfaction of users, especially in situations where user expectations are high and requirements are relatively flexible. For example, when looking for a restaurant for lunch as shown in Figure 1, the user only gives simple requirements such as *area, food type, and price range*. As there would be many restaurants satisfying such requirements, the agent can easily return one restaurant name and end with good task completion scores. However, such a response might hardly satisfy the user. For one thing, the user might not be familiar with the recommended ones, thus providing more information in the response is necessary (such as the 4th response). And for another, with limited requirements explicitly expressed, there would be latent expectations of the user. Generally speaking, in real-world cases, inquiring of human agents might receive additional information based on his/her personal experience and perceptions. It has the potential to activate certain interests of the user reciprocally. Such active sharing as in the 5th response would make the response more vivid and enhance the user experience. Nonetheless, current research on response generation for task-oriented dialogues has largely ignored these.

In this work, we target at analyzing the correlations between various agent behaviors and user satisfaction levels. Based on the findings from our analysis, we then design a neural model **EnRG** for *En*-riched *R*-esponse *G*-eneration. It incorporates crowd-sourced knowledge into responses for gaining higher user satisfaction beyond pure task completion. It adaptively decides what actions to take and which modality to express at semantic dialogue act level. Specifically, given the dialogue history, we use a single causal language model trained on both dialogue act prediction and response generation. It takes advantage of the superior performance of pre-trained GPT-2, thus possesses strong prior knowledge on how to compose fluent and coherent sentences. During the response generation stage, two special gates explicitly control the generation process. The copy gate activates the dual attention mechanism to selectively copy text from crowd-sourced knowledge for response enrichment, while the image gate encourages ranking of image candidates from it as part of the response. Extensive experiments show that while achieving comparable task completion performance, EnRG manages to generate responses gaining higher user satisfaction.

Our contributions are summarized as follows:

- We propose to consider user satisfaction beyond task completion when generating responses, pointing out the need of modeling various agent behaviors.
- We present a neural response generation method that maintains the strength of pre-trained model and makes good use of crowd-sourced knowledge via dual attention and gating mechanism.
- Extensive experiments support our concern on the limitation of current response generation in task-oriented dialogues. Both automatic and human evaluation results demonstrate the effectiveness of our proposed model.

In what follows, we will first discuss some related works and then elaborate on the correlation analysis we carried out on a public dataset. Based on the findings, we design a neural model towards enriched response generation in Section 4, followed by the experimental results in Section 5.

2 RELATED WORK

2.1 Diversifying Response in Chit-chats

Learning to generate enriched responses has long been a target to achieve for the dialogue research community. In an open-domain dialogue system, a popular model for response generation is the sequence-to-sequence (Seq2Seq) model [30]. However, the vanilla Seq2Seq models tend to provide generic responses such as "I don't know", "I am sorry", etc. Therefore, Li et al. [13] argued that using Maximum Mutual Information as the objective function can generate more diversified and relevant responses. From another aspect, by conditioning responses on both conversation history and external "fact", Ghazvininejad et al. [6] generalized the Seq2Seq approach to allow the model to produce more confluent responses without slot filling. Some efforts adopted generative adversarial networks [14, 38]. They usually formulated a reinforcement learning objective and introduced heuristics into rewards to encourage forward-looking, interactive, and informative responses. More recently, Jiang et al. [9] proposed a frequency-aware cross-entropy loss function to improve the diversity of generation models. Despite their success, these efforts focused on open-domain dialogue and cannot be directly applied to task-oriented settings. They cannot guarantee to preserve the semantics of responses which is an essential requirement for task-oriented systems.

2.2 Task-oriented Response Generation

Task-oriented dialogue system is typically composed of Natural Language Understanding (NLU), Dialogue State Tracking (DST), Policy Learning (PL), and Natural Language Generation (NLG). Response generation is closely related to Natural Language Generation (NLG) techniques, which range from simple template-based systems to machine-learned systems. The first generation of automatic NLG uses rule-based or data-driven pipeline methods. Later we witness a paradigm shift towards learning representations from large textual corpora in an unsupervised manner using deep neural network models. In the case of the task-oriented dialogue system, Kale and Rastogi [10] proposed to combine templates with Transformer models to produce utterances. Lippe et al. [18] leveraged simple templates and paraphrasing techniques to improve the diversity of generated responses. To further incorporate knowledge, some proposed to integrate knowledge bases into dialogue via memory network [20, 37]. Similarly, to translate table information to text, Chen et al. [3] introduced a content selection mechanism into neural language models and managed the two modules by a soft-gated mechanism.

To guarantee the semantics of responses, several methods proposed to predict dialogue acts first and then generate responses [21, 25]. They modeled dialogue acts in different ways, such as one-hot vectors [1, 35], multi-layer graph [2] or text sequences. For instance, Wang et al. [32] modeled the hierarchical structure between dialogue policy and NLG with a reinforcement learning framework. Chen et al. [2] proposed a pipeline to learn the dialogue act graph first and then augment the information to the neural response generation models. On the other hand, Wang et al. [34] argued that co-training the dialogue act prediction and response generation can further boost the model's performance. However,

the above-mentioned efforts all focus on task completion and surface realization of responses. User satisfaction beyond task completion is largely ignored. We instead try to enrich the response generation model with crowd-sourced knowledge considering the effect of various agent behaviors.

2.3 Large Scale Pre-Trained Models

Many of the current best-performing methods for various NLP tasks adopt a combination of pre-training followed by supervised fine-tuning using task-specific data. Pre-trained models has also been widely used in conversational systems due to their strong learning capabilities. Starting from word embeddings [22, 26], sentence embeddings [11], we are now equipped with popular pre-trained language models such as BERT [4] on classification tasks and GPT-2 [27] on generation tasks. Such models are pre-trained on large-scale open-domain corpora, and provide down-streaming tasks with rich prior knowledge while boosting their performance. For example, pre-trained on 147M conversations, DialoGPT [39] generates more relevant, confluent and context-consistent responses than strong baseline systems. There are also efforts adopted pre-trained language models like GPT-2 to boost their performance in dialogue response generation [3, 8]. In this paper, we also take advantage of the pre-trained language model GPT-2 to enhance the quality of generated responses. In addition, we look further to leverage external knowledge to enrich the generation process and make the generated responses richer in content.

3 CORRELATION ANALYSIS

We use the open-source MMConv dataset [15] to analyze the correlations between various agent behaviors and user satisfaction. It is a multi-modal dialogue corpus spanning multiple domains, where a crowd-sourced knowledge database storing user reviews and images is present together with the conversations. We focus on the restaurant domain here as the multi-domain applicability is not the focus here. In general, we emphasize three kinds of agent behaviors of using the crowd-sourced knowledge which we will elaborate in more details later.

3.1 Dataset Background

Different from the datasets collected in the Wizard-of-Oz framework (WOZ) [1, 5] that suffer from the hardship of ensuring dialogue coherence, the MMConv dataset [15] consists of human-generated multi-modal conversations, where external user reviews and photos are incorporated into the system responses. Each conversation is fully annotated with dialogue belief states and dialogue acts. We calculate the statistics for the sub-sampled dialogues as shown in table 1, where the "user goals" stand for the set of target venues with some details the user is looking for.

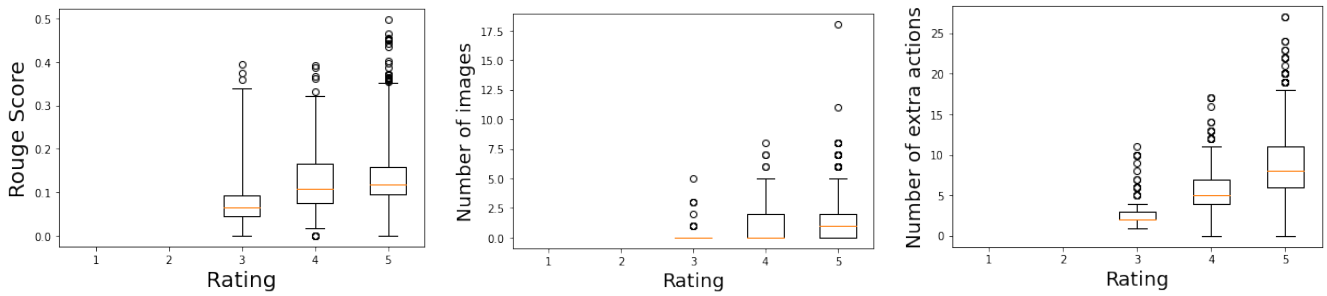
In the dataset, each dialogue contains a user feedback rating score ranging from three to five to reflect the satisfactory level. The five refers to successful, responsive, informative responses that satisfy users the most, while three means basic task completion criteria is met. The dataset also contains a crowd-sourced knowledge base from Foursquare City Guide about a famous tourist city, Singapore. The database consists of 1, 164 restaurants with 27, 618 user reviews (tips) and 67, 554 photos. Reviews for each venue often offer detailed feedback based on user experiences. They are ranked in descending order by the number of up-votes provided by other social users. More specifically, information about each restaurant is stored in a dictionary, where user views are in the $Tip\ x : < sentence >$ format and x stands for the rank. In addition, other basic attributes such as address, price, category of food *etc.* are also available [15]. Some examples of crowd-sourced knowledge are shown in Table 2.

Table 1: The general statistics of the MMConv dialogues on restaurant domain.

Entry	Number
# dialogues	1,491
# dialogue turns	11,408
# single v.s. multi-modality	282 v.s. 1,209
# unique user goals	110
# total venues in DB	1,164
# total images in DB	67,554
# total reviews in DB	27,618
# average user ratings	4.67

Table 2: Crowd-sourced knowledge examples. The information is presented non-exclusively. Contact numbers are masked for privacy issue.

Key	Value	Vote
Name	Eng Seng Restaurant	
Contact	+65 **** **67	
Region	Bedok	
Price	Expensive	
Type	SeaFood	
Tip 1	Extremely delicious! If you're just two persons, don't bother to order the other dishes, black pepper crab it is!	2
Tip 2	Crab is in the house and they are really fantastic: juicy and buttery. It's a beautiful mess to eat them and it's one of the few foods where it takes longer to eat than to cook it...	1
Tip 3	All times favourite! The best black pepper crab i tried so far! Can't wait to go back again! Awesome! *Go early to avoid long queue!	1
Name	Nana's Green Tea	
Contact	+65 **** **12	
Region	Central Region	
Price	N.A.	
Type	Cafe	
Tip 1	Surprisingly good food and desserts. Attentive staff with great service. Highly recommended. Prices are fair too.	6
Tip 2	Their weekday lunch sets are so worth it! Main and a drink for S$16.40 taxes and service included!	2
Tip 3	The matcha anmitsu is a great dessert!	2



(a) Leveraging user tips (b) Making use of images (c) Performing extra actions

Figure 2: Correlations between user rating score and agent's behaviors from dialogues on the restaurant domain. The ROUGE score in the left image is calculated by measuring the overlap between agent's response and customer reviews of the target venue.

3.2 Findings

To investigate the correlations between agent behaviors and user satisfaction scores, we center on three kinds of agent behaviors here: (a) include user reviews, (b) add image responses, and (c) execute extra actions. The first one is natural and intuitive. The user reviews are often commented based on existing user experiences, hence would offer detailed information such as popular dishes or dining environment. If the agent can make good use of such information, the richness of responses would be greatly boosted. The second one is also straightforward. The images of restaurants in the database come from professional or smartphone cameras shooting in location, which serves as a convenient and vivid method of expression. Therefore, adding images to responses also has the potential to improve user satisfaction. Last but not the least, we measure the activeness of the agent via the number of extra actions besides answering the user's explicit requests. For instance, if the agent informs extra suggestions beyond answering users' questions, the agent is considered relatively active. Such behaviors might also enhance the user experience.

We use box plots to demonstrate the correlations between user rating scores and the above-mentioned agent behaviors for dialogues on the restaurant domain in Figure 2. Specifically, Figure 2 (a) shows a positive correlation between user rating and the amount of tips used. Figure 2 (b) presents a positive correlation between user rating and the number of images involved in responses. Note that the image numbers are calculated at the dialogue level. Moreover, we plot the correlation between user rating and agent's initiative in Figure 2 (c). It shows that besides answering user's requests, the number of new dialogue acts of agents also affects user ratings.

4 THE ENRG METHOD

Based on the findings, we propose a neural model **EnRG** towards *En*-riched *R*-esponse *G*-eneration as shown in Figure 3. It naturally combines the power of the pre-trained GPT-2 model and two gating mechanisms in incorporating the abundant crowd-sourced knowledge in various forms.

To facilitate model description in detail, we first introduce our task formulation here. Let $C_k = \{U_1, R_1, \cdots, U_{k-1}, R_{k-1}, U_k\}$ denote the dialogue history at turn k in a multi-turn conversational setting, where U_i and R_i are the i-th turn user and system utterance, i.e., we consider the entire dialogue history for the input during training and inference. Images in dialogue history are transformed into textual class labels such as "beef", "pasta" "restaurant", etc. and concatenated to corresponding turns. Belief state sequences $B_k = \{b_1^a, b_1^s, b_1^v, \cdots, b_i^a, b_i^s, b_i^v\}$ are prepared by flattening the belief states at each dialogue turn, where each element of the sequences $\{b_i^a, b_i^s, b_i^v\}$ contains a triplet of action, slot, and value, e.g., "*inform, price, expensive*" [15]. $D_k = \{d_1, d_2, \cdots, d_l\}$ includes the tip candidates of venues close to target for current turn. Dialogue act sequences $A_k = \{a_1^a, a_1^s, a_1^v, \cdots, a_j^a, a_j^s, a_j^v\}$ have a similar structure as the belief state sequences while it denotes the actions of system instead. Our objective is to generate the dialogue acts A_k, correspondingly a natural language response $R_k = \{y_1, y_2, \cdots, y_m\}$ of m words, and probably an image response I_k from image repository. Examples of the belief state and dialogue act sequences can be found in Table 3.

Table 3: Examples of belief state and dialogue act Sequences in multiple turns

Term	Content
C1	<user> hi, i am looking for somewhere to have some thai food for dinner. could you help me find somewhere to go to?
B1	inform menu dinner; inform open-span thai food
A1	request open-span kind of place
R1	sure, can you tell me more about the kind of place you're looking for?
C2	<user> hi, i am looking for somewhere to have some thai food for dinner. could you help me find somewhere to go to? <agent> sure, can you tell me more about the kind of place you're looking for? <user> i am looking for somewhere that also serves thai iced tea.
B2	inform menu dinner; inform open-span thai food; inform open-span iced tea
A2	request open-span anything else
R2	is there anything else you would like?

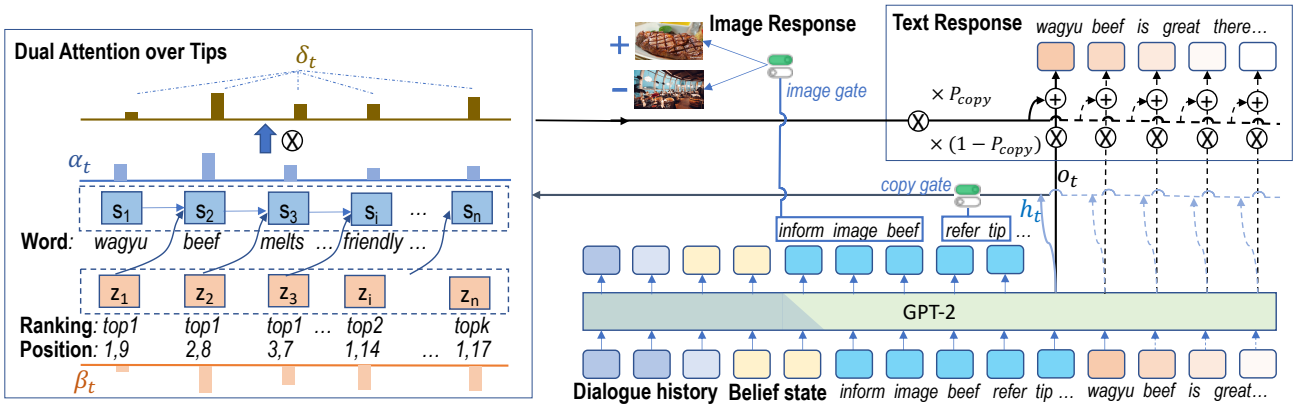


Figure 3: Illustration for EnRG model. The pre-trained language model is seamlessly incorporated with two gating mechanism to introduce crowd-sourced knowledge in response generation. The GPT-2 tokenization is ignored for ease of illustration.

4.1 Basic Generator via Pre-trained LM

We first introduce the basic generator based on pre-trained language model GPT-2. It is responsible for two tasks: generating dialogue act sequences and textual response. For each dialogue turn k, the full training sequence is composed of $X_k = [C_k; B_k; A_k; R_k]$, which is similar to SimpleTOD [8]. In general, given the dialogue history C_k and belief state B_k, the generator will be trained to generate dialogue act sequence A_k and text response R_k sequentially. Formally, the dialogue act sequence A_k is generated by:

$$A_k = \text{GPT-2}([C_k, B_k]).$$

The response R_k is then generated following the generated dialogue act sequence as

$$R_k = \text{GPT-2}([C_k, B_k, A_k]).$$

The goal of language modeling is to learn the joint probability $p(X_k)$. Suppose we view X_k as a sequence of tokens $x_1, x_2, \cdots, x_{L_k}$, where each x_i comes from a fixed set of symbols and L_k is the length. The probability can be naturally factorized using the chain rule as

$$p(X_k) = \prod_{i=1}^{L_k} p(x_i | x_{<i}).$$

Therefore, the loss for training the generator (a neural network with parameters θ) is the negative log-likelihood over a dataset composed of $|K|$ training instances. In the language model, we aim to minimize the loss as below:

$$L_{gen} = -\sum_{k=1}^{|K|} \sum_{i=1}^{L_k} \log p_\theta(x_i | x_{<i}) \quad (1)$$

4.2 Generator with Copy Gate

Beyond the basic generator, we aim to enrich the responses with crowd-sourced knowledge content. In the MMConv dataset, resources in the knowledge base are in two modalities, textual and visual [15]. Correspondingly, we devise two gates to explicitly control their usage based on special tokens in generated dialogue act: the copy gate guides the textual response to integrate the pre-trained language model with dual attention selection over external tips, while the image gate designates the image ranker to pick image candidates as part of the response.

We first introduce the generator with copy gate for leveraging textual crowd-sourced knowledge. Basically, an LSTM encoder encodes customer tips including not only all the surface tokens but also the ranking positions of these tips. While the decoder decodes textual response with copy mechanism via dual attention. The copy mechanism is only activated when the copy gate is switched on (*i.e.* the generated dialogue acts contain "refer tip"), otherwise the response comes from the basic generator. The image gate is designed with the same logic but different special tokens.

4.2.1 Textual Tips Encoder. In the MMConv dataset, the customer reviews for any venue in the crowd-sourced knowledge base are retrieved by their specific name of key $Tip\ x$, where x means the order of that particular review based on the number of up-votes [15]. To guarantee the quality of reviews to be copied, we only consider the top 3 reviews for each venue. Furthermore, we capture such ranking in our textual tips encoder to differentiate high-quality tips from low-quality tips. Specifically, we apply an LSTM-based neural network [7] to encode each word segment d_j in tips together with its ranking position embedding z_j into the hidden state s_j. Inspired by [12, 19], we represent the ranking position embedding for a word segment as the concatenation of tip ranking f_j, its in-order position p_j^+ and reverse-order position p_j^- in the tip sequence. Therefore, we have

$$z_j = [f_j; p_j^+; p_j^-].$$

The mechanism is illustrated in the left part of Figure 3. For example, the ranking position embedding for the first token z_1 will be $[1, 1, 9]$, indicating the top 1 tip, in-order position 1 and reverse-order position 9.

4.2.2 Decoder with Copy. At each decoding step t, a dual-attention weight δ_t is calculated to model the relevance between the decoder hidden states h_t from GPT-2 and the encoder states s and z of tips. Inspired from [19], we calculate semantic-level attention weights α_t to model the relevance between h_t and the encoded states s, and the position-level attention weights to model the relevance of h_t

and ranking position embedding z:

$$\alpha_{t_i} = \frac{e^{g(h_t, s_i)}}{\sum_{j=1}^{n} e^{g(h_t, s_j)}},\ \beta_{t_i} = \frac{e^{g(h_t, z_i)}}{\sum_{j=1}^{n} e^{g(h_t, z_j)}},$$

where $g(a, b) = \tanh(W_1 a) \odot \tanh(W_2 b)$ measures the relevance score between a and b with trainable parameters W_1, W_2.

The final attention weights δ_t considering both the word segment level and position level relevance is the normalized element-wise product of α_t and β_t:

$$\delta_{t_i} = \frac{\alpha_{t_i}.\beta_{t_i}}{\sum_{j=1}^{n} \alpha_{t_j}.\beta_{t_j}},\ \hat{s}_t = \sum_{i=1}^{n} \delta_{t_i} s_i,$$

where $\hat{s}_t$ is the final weighted sum of tips encoder hidden states. Based on that, a soft switch policy p_{copy} is maintained to choose between generating word from the basic GPT-2 decoder or copying from tips:

$$p_{copy} = \text{sigmoid}(W_s \hat{s}_t + W_h h_t + W_y y_t + b),$$

where y_t represents the decoder input at time step t. W_s, W_h, W_y, and b are all trainable parameters.

The soft pointer generator learns to alternate between copying from tips and generating based on training data. However, the discrepancy between the LSTM encoder and the GPT-2 decoder adds to the difficulty of training. Moreover, the training dialogue data is limited in scale while the tips repository is large. We thus need to explicitly "teach" the model where to copy and where to generate. Therefore, to provide the model with accurate guidance of the switching behavior, we mark the positions of tip segments in the target text. At these positions, we maximize the copy probability p_{copy} via an additional loss term. The loss function is

$$L_{gen} = L_{gen} + \lambda \sum_{d_m \in \{D_k\}} (1 - p_{copy}^m), \tag{2}$$

where L_{gen} is the original loss between model outputs and target texts. d_m means a matched token between response and tips. λ is hyper parameter as the weight for the copy loss term.

4.3 Ranker with Image Gate

Similar to the copy gate, the image gate is only switched on when "*inform image XX*" is generated in dialogue act sequence. Otherwise, the response will contain no image. For the ranker, we use the specific concept "*XX*" produced in the special token as the input sequence S. A pre-trained BERT model is used to encode S and the $[CLS]$ position output is used as query

$$q = BERT(S).$$

We then adopt the triplet ranking loss for training. Therefore, for each training instance, we randomly sample a negative image I_- for the target venue to pair with the correct image I_+. A fine-tuned EfficientNet [31] extracts concept level feature vector v for each image candidate. The objective for training is to minimize the triplet loss as below,

$$L_{ranker} = max(0, \epsilon + d(W_q q, W_v v_{I_+}) - d(W_q q, W_v v_{I_-})),$$

where ϵ is a margin and $d(\cdot)$ is the Euclidean distance function. W_q and W_v are trainable weight matrices. Intuitively, it encourages the distance between q and the negative image to be larger than that between q and the positive image by a margin ϵ.

Since the image ranker is only loosely connected to the generation model, we train it separately to make the training of the generation part more focused. Note that the generated dialogue act sequence A_k might contain multiple dialogue acts containing "inform image". For such cases, we treat each dialogue act like these as a separate instance.

5 EXPERIMENTS

5.1 Settings

5.1.1 Training Setup. The implementation is on a single Tesla V100 GPU with a batch size of 8. The input to the model is tokenized with pretrained BPE codes [29] associated with DistilGPT2 [28]. Experiments for the proposed EnRG model use default hyper parameters for GPT-2 and DistilGPT2 in Huggingface Transformers [36]. Text sequences longer than 1024 tokens are truncated. We fine-tune the EfficientNet [31] on the image repository. Images in dialogue history are mapped to textual labels and append to corresponding turn utterances. Note that GPT-2 is fine-tuned from pre-trained parameters while the LSTM encoder and attention part is learned from scratch, the initial geometry of the two are different. Therefore we apply larger weight to the copy loss to 'teach' the model to learn faster and better.

5.1.2 Evaluation Metrics. Following existing response generation works, we adapt the *Success Rate* and the *Inform Rate* to measure task completion – whether the system has provided the correct entity (*Inform Rate*) and provide correct answers for attributes asked by user (*Success Rate*). The fluency of the generated response is measured by *BLEU* [24] score. The combined *Score* for action and response generation is computed as $(BLEU + 0.5 \times (Inform + Success))$ [1, 34]. For image response evaluation, we use the *Recall@m* metric similar to [16, 23]. The result is correct only if the positive image is ranked in the top-m ones. Moreover, we use micro *Inform F1* and *Request F1* to evaluate the *inform* and *request* actions for dialogue act prediction.

Since these metrics are largely confined in task completion or surface fluency, we further carry out human evaluation to measure the user satisfaction levels. We add three criteria as follows: *Informativeness* measures if the response provides relevant and useful information to user; *Impressiveness* reflects how imposing and rich the response is; *Involvingness* shows how active and involving the agent is. These are closely related to the agent behavior patterns we investigated.

5.1.3 Baselines.

- **LSTM** [1]: It frames the dialogue as a context to response mapping problem, a sequence-to-sequence model. The LSTM base module is applied.
- **Mem2Seq** [20]: It loads entities in KB and dialogue history into memory. During the decoding stage, it uses a gate to decide whether to copy an entry from memory or generate a word from the vocabulary.
- **DialoGPT** [39]: DialoGPT is based on the structure of GPT-2 [27]. It is pre-trained on large-scale Reddit conversation-like comments.

- **MARCO** [33]: It considers dialogue act prediction as a sequence generation problem, and uses a model to jointly generate act and response.
- **SimpleTOD** [8]: It uses a large pretrained language model, GPT-2 [27] to solve multiple tasks in task-oriented dialogues. It reads dialogue history as context, performs dialogue state tracking, dialogue act prediction followed by response generation.

5.2 Automatic Evaluation

5.2.1 Main Results. The main results for response generation are presented in Table 4, where *Score* means the combined score. From the table, we observed that the proposed EnRG method outperforms all baselines in most metrics. First of all, we observe that transformer-based methods like DialoGPT, SimpleTOD, and EnRG outperform LSTM network-based methods such as basic LSTM and Mem2Seq by relatively large margins. This is mainly attributed to the relatively better learning capabilities of transformer models. Secondly, modeling responses in semantic dialogue act level boosts performance for response generation. This is evidenced by the better performance of SimpleTOD and EnRG as compared to that of DialoGPT, which does not consider belief states and dialogue acts. The performance gap is especially evident in the task completion metrics such as *Inform rate*. Thirdly, the usage of crowd-sourced knowledge via copy has a positive effect on improving the BLEU score for EnRG as demonstrated by the 4.2% performance improvement over SimpleTOD. It shows that the copy mechanism in EnRG can incorporate appropriate knowledge while preserving the comprehensiveness of the responses. To further evaluate the informativeness of EnRG-generated responses, we compared it with the state-of-the-art SimpleTOD via human evaluation.

Table 4: The main results for responses.

Method	Inform	Success	BLEU	Score
SC-LSTM	2.19	1.73	16.47	18.43
Mem2Seq	55.94	24.48	11.21	51.42
DialoGPT	64.50	47.56	19.92	75.95
MARCO	77.78	47.22	18.63	81.13
SimpleTOD	**80.97**	47.74	22.73	87.09
EnRG	78.22	**49.66**	**23.68**	**87.62**

5.2.2 Results on Dialogue Acts. Besides the main results, we compare the performance on agent dialogue act prediction in Table 5. We evaluate it as a multi-label classification problem that considers action and slot pairs such as "inform-outdoor seating". Note that although MACRO also models dialogue acts, it separates actions and slots into groups respectively, thus we cannot generate an F1 score for the pairs. From Table 5, we observe that EnRG yields better *Inform F1* score which is consistent with the winning result of EnRG in *Success rate* in Table 4. It shows that EnRG manages to provide information correctly.

We further test the performance of the proposed EnRG on controlling gates. In general, the high scores in Table 6 demonstrate that EnRG works well. For example, the F1 score for the copy gate exceeds 72% while that for the image gate exceeds 80%. When inspecting the responses, we found that this is mainly due to the obvious signals in dialogues. For example, the agent usually makes use of user tips when recommends a restaurant to the user or the user explicitly requests tips or suggestions. Similarly, the agent often provides images when recommends restaurants or the user requests for details of food or dining.

Table 5: Dialogue acts prediction results.

Method	Inform F1	Request F1
SimpleTOD	50.29	18.80
EnRG	**50.98**	**19.96**

Table 6: EnRG's performance on gate prediction.

Metrics	Copy Gate	Image Gate
Precision	65.8	89.6
Recall	81.4	77.0
F1 Score	72.8	82.8

5.3 Human Evaluation

5.3.1 Evaluation Settings. We conduct a human study to evaluate our model by the AMT crowd-sourcing platform. For this purpose, we randomly selected 100 sample dialogues from the test dataset and constructed two groups of systems for comparison: Ours vs. SimpleTOD and Ours vs. Human Response, where Human Response means the reference responses. Responses generated by each group were randomly assigned in pairs to 3 judges, who ranked them according to *Informativeness*, *Impressiveness* and *Involvingness*.

5.3.2 Results. The results of this study are shown in Figure 4. Overall speaking, EnRG outperforms SimpleTOD and is relatively comparable to Human Response. First of all, we observe many tie cases for EnRG and SimpleTOD. This is because these two methods rely on the same base network and formulate the task in a similar style. Some of their generated responses are even identical to each other. Nonetheless, EnRG outperforms SimpleTOD in *Informativeness* and *Impressiveness*, and exceeds it by 16.0% and 13.2% of wining cases respectively. This is as expected because EnRG explicitly leverages extra crowd-sourced tips to make the generated responses more informative, while also provides image response which is a rather convenient way of expression. For the *Involvingness* metric, EnRG is relatively weak. However, the Success Rate in Table 4 shows that EnRG answers requests better for recommended venues. This might be because the definition of *Involvingness* is relatively subjective thus affected annotators' judge.

When compared to Human Response, we observe fewer tie cases while the number of win and lose cases are relatively closely contested. Specifically, there are fewer than 1/3 of testing cases with tie results across all three measures. It shows a relatively large variance between human-generated responses and machine-generated ones. This is reasonable as the Human responses are generated by 87 different human annotators while the machine targets at learning the most salient patterns from it. Regarding *Informativeness*, EnRG wins for 31% of times while loses for 40% of times. It shows that EnRG generated responses are indeed informative and rich. For *Impressiveness*, the distance between them is even narrower. Through detailed analysis of testing cases, we observe that EnRG tends to

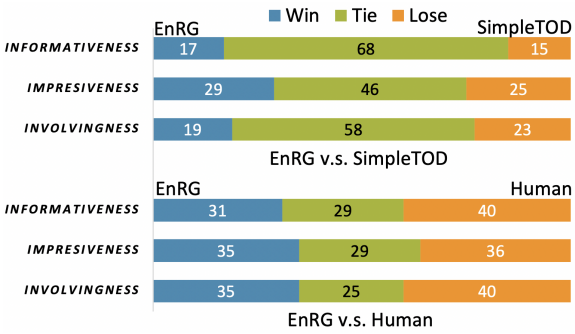


Figure 4: Human study results in response quality. "Win", "Tie" or "Lose" indicate the proportions that the EnRG wins over, ties with or loses to its counterpart respectively.

provide extra image responses for these recommendation turns, especially for these testing dialogues with lower user ratings. This signals the importance and potential of introducing visual modality into task-oriented dialogue. Together with the results in Table 4, it is clear that although EnRG obtains similar results to the strongest baseline SimpleTOD for existing popular metrics, it manages to achieve better user satisfaction and is close to human responses in these aspects.

5.4 Case Study

In order to better illustrate the difference between our proposed EnRG and the SimpleTOD model, we also do case study on the generated responses. Some examples are shown in Table 7.

Generally speaking, we observe that the proposed EnRG method is able to make good recommendations while also providing useful information. For example, for the dialogue 5000 (the first example in Table 7), both the EnRG generated response and the response in dataset recommend venues while the baseline SimpleTOD generated response chooses to ask more questions about "local food". Besides, we observe that the EnRG generated response contains more useful information selected from crowd-sourced reviews, such as "authentic malay cuisine", "authentic indian food" and "great value". These are usually important information for users to consider when making their decisions. That is to say, EnRG tends to provide useful information about the venues to make the recommendation more convincing.

A similar case also happens in dialogue 6236. When the user requests for a cafe to have a rest and providing some extra information in the context such as "great value" and "central region", the EnRG method generates venue recommendation and offers tips about "dim sum" which corresponds to the "scones" in the dialogue context. The SimpleTOD method only manages to generate a recommendation and the response tends to be more general than the reference response in the MMConv dataset. Note that dialogue 6236 receives a rating score of four, so the response from the dataset is generally preferred by the users. We observe that it indeed contains extra information such as other food images.

These examples altogether demonstrate that responses from EnRG might be more content-rich than the responses generated by baseline methods or even the responses in the dataset in certain circumstances. Such attributes of EnRG demonstrate the design of incorporating crowd-sourced knowledge and the effectiveness of the proposed model in realizing such incorporation.

Table 7: Examples of responses generated by EnRG, SimpleTOD, and the corresponding responses from the MMConv dataset. The exact venue names, addresses, etc. are replaced with placeholders for delexicalization.

ID	5000
Rating	5
Context	<user> i will like a place for a quick meal. <agent> what location will you like? <user> i will like bishan.
EnRG	i recommend [venuename]. they sell authentic malay cuisine and serve authentic indian food at great value.
SimpleTOD	will you like some local food?
Dataset	i recommend [venuename].
ID	6236
Rating	4
Context	<user> hi, i would like to have a rest in a cafe, any recommendations? <agent> sure, would you like to try scones like the one in the image? <user> looks great, can you recommend a cafe that serves this food with great value in central region?
EnRG	you can go to [venuename]! it is a great place for dim sum.
SimpleTOD	you can visit [venuename]!
Dataset	you can try [venuename]! these are other food images from the cafe!

6 CONCLUSION

In this paper, we explored various factors for generating enriched responses in task-oriented dialogues under the help of crowd-sourced knowledge. The ultimate goal is to improve user satisfaction without impairing task completion. We carried out detailed analysis on a fully annotated multimodal dialogue dataset and obtained several findings on user behaviors. Inspired by our findings, we proposed a neural response generation model EnRG that possesses the superior performance of pre-trained GPT-2 while effectively leverages external knowledge in both textual and visual modalities. We conducted extensive experiments and the results demonstrated the effectiveness of our proposed model. Our work would shed lights on the current limitation of evaluation metrics for task-oriented responses and the potential of introducing multi-modality. In future, we will further explore more advanced methods for better incorporating visual modality into task-oriented dialogue systems.

ACKNOWLEDGMENT

This research is supported by the Sea-NExT Joint Lab. The authors thank the anonymous reviewers for their valuable comments, which are beneficial to the authors' thoughts on the response generation task and the revision of the paper.

REFERENCES

[1] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *EMNLP*. 5016–5026.

[2] Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically Conditioned Dialog Response Generation via Hierarchical Disentangled Self-Attention. In *ACL*. 3696–3709.

[3] Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2020. Few-Shot NLG with Pre-Trained Language Model. In *ACL*. 183–190.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[5] Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. arXiv:1907.01669 [cs.CL]

[6] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932* (2017).

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[8] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A Simple Language Model for Task-Oriented Dialogue. In *NeurIPS*. 1–13.

[9] Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving Neural Response Diversity with Frequency-Aware Cross-Entropy Loss. *WWW* (2019).

[10] Mihir Kale and Abhinav Rastogi. 2020. Template Guided Text Generation for Task-Oriented Dialogue. In *EMNLP*. 6505–6520.

[11] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

[12] Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural Text Generation from Structured Data with Application to the Biography Domain. In *EMNLP*. 1203–1213.

[13] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. arXiv:1510.03055 [cs.CL]

[14] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep Reinforcement Learning for Dialogue Generation. arXiv:1606.01541 [cs.CL]

[15] Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. MMConv: An Environment for Multimodal Conversational Search across Multiple Domains. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[16] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*. 801–809.

[17] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[18] Phillip Lippe, Pengjie Ren, Hinda Haned, Bart Voorn, and Maarten de Rijke. 2020. Diversifying Task-oriented Dialogue Response Generation with Prototype Guided Paraphrasing. *arXiv preprint arXiv:2008.03391* (2020).

[19] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *AAAI*. 4881–4888.

[20] Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems. In *ACL*. 1468–1478.

[21] Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured Fusion Networks for Dialog. In *SIGdial*. 165–177.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*. 3111–3119.

[23] Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal dialog system: Generating responses via adaptive decoders. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1098–1106.

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*. 311–318.

[25] Jiahuan Pei, Pengjie Ren, and Maarten de Rijke. 2019. A modular task-oriented dialogue system using a neural mixture-of-experts. *arXiv preprint arXiv:1907.05346* (2019).

[26] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. 2227–2237.

[27] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog* 1, 8 (2019), 9.

[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[29] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL*. 1715–1725.

[30] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215 [cs.CL]

[31] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *ICML*. 6105–6114.

[32] Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2020. Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system. *arXiv preprint arXiv:2006.06814* (2020).

[33] Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. 2020. Multi-Domain Dialogue Acts and Response Co-Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7125–7134.

[34] Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. 2020. Multi-Domain Dialogue Acts and Response Co-Generation. *arXiv preprint arXiv:2004.12363* (2020).

[35] TH Wen, M Gašić, N Mrkšić, PH Su, D Vandyke, and S Young. 2015. Semantically conditioned lstm-based Natural language generation for spoken dialogue systems. In *EMNLP*. 1711–1721.

[36] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*. 38–45.

[37] Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2018. Global-to-local Memory Pointer Networks for Task-Oriented Dialogue. In *ICLR*.

[38] Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In *EMNLP*. 3940–3949.

[39] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv* (2019), arXiv–1911.