Topic-Aware Response Generation in Task-Oriented Dialogue with Unstructured Knowledge Access

Anonymous ACL submission

Abstract

To alleviate the problem of structured databases' limited coverage, recent task-oriented dialogue systems incorporate external unstructured knowledge to guide the generation of system responses. However, these usually use word or sentence level similarities to detect the relevant knowledge context, which only partially captures the topical level relevance. In this paper, we examine how to better integrate topical information in knowledge grounded task-oriented dialogue and propose "Topic-Aware Response Generation" (TARG), an end-to-end response generation model. TARG incorporates multiple topic-aware attention mechanisms to derive the importance weighting scheme over dialogue utterances and external knowledge sources towards a better understanding of the dialogue history. Experimental results indicate that TARG achieves state-of-the-art performance in knowledge selection and response generation, outperforming previous state-of-the-art by 3.2, 3.6, and 4.2 points in EM, F1 and BLEU-4 respectively on Doc2Dial, and performing comparably with previous work on DSTC9; both being knowledge-grounded task-oriented dialogue datasets.[1]

1 Introduction

Task-oriented (or goal-oriented) dialogue systems aim to accomplish a particular task (e.g. book a table, provide information) through natural language conversation with a user. The system's available actions are often described by a pre-defined domain-specific schema while relevant knowledge is retrieved from stuctured databases or APIs (Rastogi et al., 2020). As such, task-oriented dialogue systems are often limited on which actions can be taken and what information can be retrieved (Kim et al., 2020). To relax these restrictions, some dialogue systems (also referred to as goal-oriented chatbots) adopt open-domain language that is by

[1]Code will be made public on the paper's acceptance.

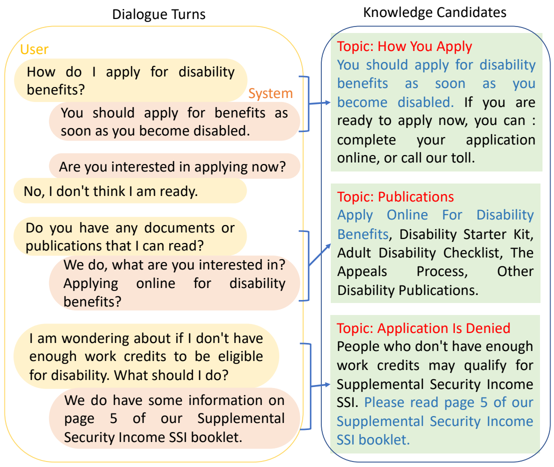

Figure 1: An example of knowledge-grounded dialogue.

definition unconstrained by pre-defined actions (Feng et al., 2020), and dynamically extract any required knowledge from in-domain unstructured collections in the form of entity descriptions, FAQs, and documents. Access to external knowledge sources has also been shown to help dialogue systems generate more specific and informative responses, which helps with the "common response" problem (Zhang et al., 2018; Ren et al., 2020).

Figure 1 shows an example of a task-oriented dialogue that exploits external unstructured knowledge sources. Given a history of previous dialogue turns, with each turn consisting of one user and system utterance, and access to in-domain unstructured knowledge sources (either a document collection or a set of candidate facts), the dialogue system needs to generate an appropriate system response for the current turn. Recent research (Zhang et al., 2018; Ren et al., 2020) tackles the task by decomposing it into two sub-tasks: to initially determine the relevant knowledge (if any) that needs to be extracted/selected from external resources, and to subsequently generate the response based on the selected knowledge and the dialogue history.

When retrieving knowledge from unstructured sources, different sources may need to be accessed in different dialogue turns; this is to be expected in most conversation scenarios. In the example of Figure 1, the first turn is grounded on the first knowledge candidate, and subsequent turns are grounded on later candidates. If we consider that each knowledge source belongs to a different topic or domain (e.g. "how you apply", "publications", "application is denied" in our example), we can observe that as the knowledge selection shifts across sources during the course of the dialogue, a corresponding shift occurs between topics. Previous work has not actively exploited this, but we posit that attending the topic shifts in the dialogue history can provide signals that help distinguish relevant from irrelevant sources for knowledge selection, and that such topical information can help the model derive an importance weighting scheme over the dialogue history for better response generation.

In this paper, we model topic shifts in selected knowledge sources to improve topic-aware knowledge selection and response generation in task-oriented dialogue, and propose "Topic-Aware Response Generation" (TARG), an end-to-end model for knowledge selection and response generation. Our approach incorporates multiple topic-aware attention mechanisms to derive the importance weighting scheme over previous utterances and knowledge sources, aiming for a better understanding of the dialogue history. In addition, TARG is built on top of recent breakthroughs in language representation learning by finetuning on the pre-trained language model BART (Lewis et al., 2020).

We conduct extensive experiments with two task-oriented dialogue datasets, namely Doc2Dial (Feng et al., 2020) and DSTC9 (Gunasekara et al., 2020). Our results indicate that TARG is able to accurately select the appropriate knowledge source, and as a result generate more relevant and fluent responses, outperforming previous state-of-the-art by 3.2, 3.6, and 4.2 points in EM, F1 and BLEU-4 respectively on Doc2Dial, and performing comparably with previous work on DSTC9. Furthermore, we present an ablation study and a case study accompanied by analysis of the learned attention mechanisms.

2 Related Work

As we briefly mentioned in the introduction, the majority of previous work decomposed knowledge-grounded dialogue generation into two sub-tasks: knowledge selection and response generation.

To determine the relevant candidate for knowledge selection, the use of keyword matching (Ghazvininejad et al., 2018), information retrieval (Young et al., 2018) and entity diffusion (Liu et al., 2018) methods have been proposed. More specifically, keyword matching methods (Bordes et al., 2017) focus on calculating a weight for each keyword in the knowledge candidate and then determine their relevance based on the weighted sum of the keywords' representations. On the other hand, some information retrieval techniques compute traditional *tf-idf* scores to detect the knowledge candidate in the most relevant document to the user's query (Song et al., 2018; Dinan et al., 2018), while others leverage the power of neural networks to learn a candidate ranking function directly through an end-to-end learning process (Yan and Zhao, 2018; Zhao et al., 2019; Gu et al., 2019, 2020). Another approach uses entity diffusion networks (Wang et al., 2020) that perform fact matching and knowledge diffusion to ground both knowledge candidates and dialogues.

For response generation, the related work has adapted both response retrieval and language generation approaches. Specifically for response retrieval, deep interaction networks (Sun et al., 2020) have been employed to learn better-suited representations to ground candidate responses against external knowledge, while language generation approaches have been adapted to attend to ground knowledge during inference (Peng et al., 2020), with some further employing copy mechanisms over both dialogue context and external knowledge (Yavuz et al., 2019), or leveraging a reading comprehension model to similarly extract relevant spans (Qin et al., 2019; Wu et al., 2021).

Recently, pre-trained language models such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), which have demonstrated significant improvements on numerous natural language processing tasks, have also been applied to improve model the semantic representation in knowledge selection and response generation (Zhao et al., 2020; Li et al., 2020; Feng et al., 2020). Alternatively, other approaches combine the generative capability of autoregressive decoders such as GPT-2 (Budzianowski and Vulić, 2019) or T5 (Raffel et al., 2020), to better generate the system response.

Broader dialogue research has explored the topic-aware signal present in the dialogue history, but such work did not consider external knowledge nor its topics. Briefly, Xing et al. (2017) proposed a

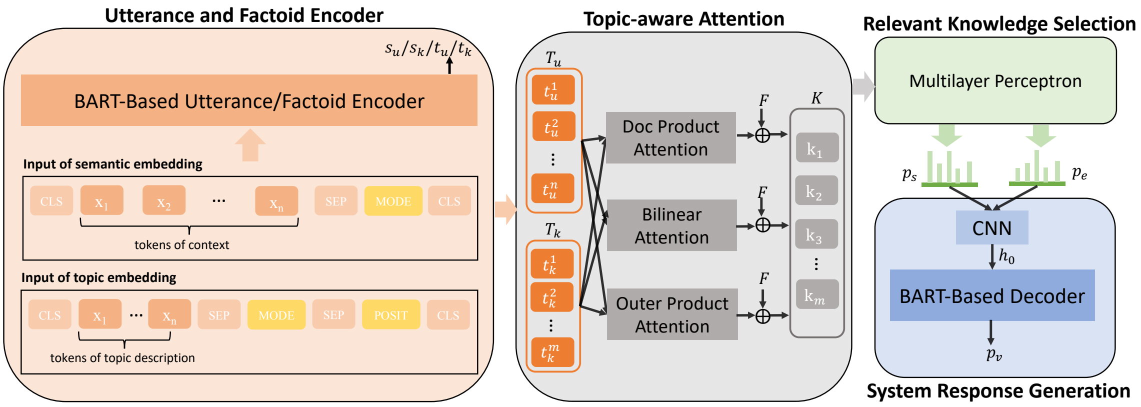

Figure 2: Overview of Topic-Aware Response Generation (TARG).

topic-aware seq-to-seq approach for open-domain dialogue that attends over LDA topics inferred from the dialogue history, while Zhang et al. (2020) calculates the relevance between topic distributions of the dialogue history and the immediate context and attends over them to generate the next system response. In retrieval-based dialogue systems, Xu et al. (2021b) performs topic-aware segmentation of the context to better inform dialogue modeling.

We briefly discuss more recent work in our experiments section, as we compare it against our approach. To the best of our knowledge no other work has explicitly modelled the topic shifts in both dialogue history and external knowledge to inform knowledge selection and response generation in knowledge-ground task-oriented dialogue systems.

3 Our Approach

As we mentioned in the introduction, our proposed approach (TARG) exploits topic-aware mechanisms to derive an importance weighting scheme over different utterances in the dialogue history, with the goal to better inform knowledge selection and response generation. For a brief overview of TARG, please consult Figure 2. The input in our task consists of the dialogue history of previous user and system utterances, and a set of external knowledge candidates (hereafter referred to as factoids for brevity). The goal is to generate the next system utterance in the dialogue, which may or may not be grounded in one of the factoids; some of the dialogue history utterances may also be grounded on factoids but not necessarily all of them are.

Briefly, to generate the next turn's system utterance, TARG initially generates BART-based representations for every previous user and system utterance in the dialogue history, for every available factoid, and for both utterances' and factoids' corresponding topics. For each utterance / factoid pair, TARG extracts matching features by calculating feature interaction over their encoded representations. TARG subsequently weights the matching features by topic-aware attention mechanisms, and aggregates them in a tensor. Finally, a knowledge selection layer outputs a relevance score over factoids, and the decoder generates the system utterance based on the most relevant factoid's encoding.

3.1 Utterance and Factoid Encoder

We use a BART encoder to generate representations for every utterance in the dialogue history (up to a maximum history length) and factoid in external knowledge. We similarly, but separately, generate representations for their corresponding topics. Our work assumes that the corresponding topic of factoids can be derived in some way from the available data, e.g. the topic can be interpreted as the title of the factoid's originating document or its annotated domain. While we do not explore the possibility in this paper, the topic could also potentially be inferred using topic modelling techniques. The topic of each utterances is considered the same as that of their corresponding factoids (if any). Since not all dialogue turns are necessarily grounded in external knowledge, in absence of a corresponding factoid, the topic is set to a generic "non-relevant" pseudo-topic. This process results in the semantics and topic of every utterance or factoid being represented explicitly by separate embeddings.

Specifically, in order to generate the semantic embeddings s_u and s_k of every utterance and factoid respectively, the token sequence $X = ([\text{CLS}], x_1, ..., x_N, [\text{SEP}], [\text{MODE}], [\text{CLS}])$ is passed through a BART encoder, where the subword tokens of the text are denoted as $x_1, ..., x_N$. [CLS] and [SEP] are start-of-text/end-of-text and separator pseudo-tokens respectively, while [MODE] is one of [SYS]/[USER]/[KLG] to indicate whether the text belongs to a system utterance, user utterance, or factoid respectively. The state of the final [CLS] is used as the utterance's / factoid's semantic embedding. Similarly, to generate the topic embeddings t_u and t_k of every utterances and factoid, the BART encoder sequence input is $T = ([\text{CLS}], x_1, ..., x_N, [\text{SEP}], [\text{MODE}], [\text{POSIT}], [\text{CLS}])$, where [POSIT] is the position of the corresponding dialogue history utterance (zero if the text belongs to a factoid). The state of the final [CLS] is used as the topic embedding.

3.2 Topic-aware Attention

In the next step, TARG calculates feature interactions over the semantic embeddings to extract matching features, which are subsequently weighted by a number of topic-aware attention mechanisms. These attention mechanisms operate over the topic embeddings of utterances and factoids to calculate topic-aware utterance / factoid pair matching representations. The motivation is to incorporate a more flexible way to weight and aggregate matching features of different dialogue history utterances with topic-aware attention, so that the model learns to better attend over them.

Specifically, we design three different types of topic-aware attention that are calculated between each topic embedding t_k^i, corresponding to the i-th factoid, and the topic embeddings of all utterances in dialogue history, as follows:

Dot Product. We concatenate the utterance topic embeddings t_u with the factoid topic embedding, and compute the dot product between parameter w_d and the resulting vector:

$$A_d^i = \text{softmax}(\exp(w_d^T[t_u, t_k^i]), \forall t_u \in T_u) \quad (1)$$

Bilinear. We compute the bilinear interaction between t_u and t_k^i and then normalize the result:

$$A_b^i = \text{softmax}(\exp(t_u W_b t_k^i), \forall t_u \in T_u) \quad (2)$$

where W_b is a bilinear interaction matrix.

Outer Product. We compute the outer product between t_u and t_k^i, then project this feature vector through a fully connected layer and a softmax:

$$A_o^i = \text{softmax}(\exp(w_o^T(t_u \times t_k^i), \forall t_u \in T_u) \quad (3)$$

where w_o is a parameter and $\times$ is the outer product.

In parallel, we calculate the feature interaction matrix F_i between the semantic embeddings of all utterances s_u^j and the factoid s_k^i. Every row $F_{i,j}$ of F_i is calculated as follows:

$$F_{i,j} = v_f^T \tanh(s_u^j W_f s_k^{iT} + b_f) \quad (4)$$

with W_f, b_f, v_f being model parameters.

To obtain a unified utterance / factoid pair representation k_i for each factoid i, we concatenate the weighted sums of all utterances / factoid interaction embeddings with the different attention mechanisms. The final topic-aware utterance / factoid pair representation across all factoids is K, where the i-th column vector is k_i:

$$k_i = [A_d^{iT} F_i, A_b^{iT} F_i, A_o^{iT} F_i] \quad (5)$$

3.3 Relevant Knowledge Selection

For the purpose of knowledge selection, TARG treats all external knowledge as a single document, by simply concatenating all available factoids. To account for the possibility that the system response shouldn't be grounded on any external knowledge, a "non-relevant" pseudo-factoid is included.

The relevant knowledge selector takes the topic-aware representations of these sequential factoids as input and predicts a span over the overall document that the system response should be grounded on. Through this process, several knowledge candidates may appear in the selected span.

The grounded span is derived by predicting the start and the end indices of the span in the document. We obtain the probability distribution of the start index and end index over the entire document by the following equations:

$$p_s = \text{softmax}(W_s^T K + b_s), \quad (6)$$

$$p_e = \text{softmax}(W_e^T K + b_e), \quad (7)$$

where W_s, W_e, b_s, b_e are trainable weight vectors.

3.4 System Response Generation

The system response generator decodes the response by attending on the selected knowledge span. Since the span may contain several factoids, we first use a Convolution Neural Network (CNN) to fuse the information. We apply this CNN even when only a single factoid is present in the span for consistency. The CNN receives the topic-aware

Domain	#Dials	#Docs	avg # per doc tk	sp	p	sec
ssa	1192	109	795	70	17	5
va	1330	138	818	70	20	9
dmv	1305	149	944	77	18	10
studentaid	966	91	1007	75	20	9
all	4793	487	888	73	18	8

Table 1: Number of dialogues, documents and average number of content elements per document (tk: tokens, sp: spans, p: paragraphs, sec: titled sections) per domain in Doc2dial.

Domain	#Dials	#Snippets	#per-snip tk	sent
Hotel	-	1219	9	1.00
Restaurant	-	1650	7	1.00
Train	-	26	15	1.20
Taxi	-	5	19	1.15
all	10,438	2900	8	1.00

Table 2: Number of dialogues, snippets and average number of content elements per snippet (tk: tokens, sent: sentences) per domain in the DSTC9 dataset.

utterance / factoid pair embeddings of the selected span, and outputs the fusion embedding:

$$f = \text{CNN}(K_{s:e,:}), \tag{8}$$

where s and e are the start and end indexes.

We employ a BART decoder for the system response generator, which takes the fusion embedding f as its initial hidden state. At each decoding step t, the decoder receives the embedding of the previous item w_{t-1}, and the previous hidden state h_{t-1}, and produces the current hidden state h_t:

$$h_t = \text{BART}(w_{t-1}, h_{t-1}). \tag{9}$$

A linear transformation layer produces the generated word distribution p_v over the vocabulary:

$$p_v = \text{softmax}(VW_v h_t + b_v), \tag{10}$$

where V is the word embeddings of the vocabulary, and W_v and b_v are transformation parameters.

3.5 Optimization

During training, we optimize both the knowledge selector and response generator via their cross-entropy losses $\mathcal{L}_s$, $\mathcal{L}_g$ respectively. We compute the joint loss $\mathcal{L}$ as follows:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_s + (1 - \lambda) \cdot \mathcal{L}_g, \tag{11}$$

where $\lambda \in [0, 1]$ is a balance coefficient.

4 Experiments

4.1 Datasets

We evaluate our proposed approach on two benchmark data sets on task-oriented dialogue: Doc2dial (Feng et al., 2020) and DSTC9 (Gunasekara et al., 2020). Doc2dial is a recently released dataset with a withheld test set used for the corresponding leaderboard, which includes conversation dialogues between an assisting system and an end user, with an accompanying set of documents wherein distinct factoids are clearly annotated; further annotations indicate which dialogue utterances are grounded on which factoids of the associated documents. The Doc2dial dataset includes many cases of conversations that are grounded on factoids from different documents. If we consider the title of each document as a distinct topic, then each of these conversations can be interpreted to involve many interconnected topics under a general inquiry, making it an ideal dataset for our approach.

The DSTC9 dataset also includes conversation dialogues, but the external knowledge is in the form of FAQ documents, in essence containing question answering pairs on a specific domain; we consider each pair as a distinct factoid and their domain as the topic. In practice, these FAQs are to be used to answer follow-up user questions that are out of the coverage of a dialogue system's database. Similarly to Doc2Dial, we observe that the focused "topic" in the DSTC9 dataset is also varied thoughout the conversations. Table 1 and Table 2 presents the statistics of the Doc2dial and DSTC9 datasets.

4.2 Baselines

In the following experiments, we compare our approach against previously published state-of-the-art approaches on the Doc2dial and DSTC9 datasets. We have not re-implemented these approaches, but report their already published results for the datasets for which they are available.[2]

Doc2Dial-baseline (Feng et al., 2020): This is the baseline provided by the Doc2Dial challenge. It consists of an extractive question answering model

[2]While there are better performing systems in the DSTC9 and Doc2Dial leaderboards, these are either not published, not based on a single method, or exploit additional external data, and thus are not directly comparable to this work.

Model	Knowledge Selection		Response Generation						
	MRR@5	Recall@5	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	0.726	0.877	0.303	0.173	0.100	0.065	0.338	0.136	0.303
KDEAK	0.853	0.896	0.355	**0.230**	0.153	0.104	0.397	**0.190**	0.357
RADGE	**0.937**	0.966	0.350	0.217	0.135	0.089	0.393	0.175	0.355
EGR	0.894	0.934	0.361	0.226	0.140	0.096	0.397	0.179	0.353
TARG	0.935	**0.972**	**0.366**	0.224	**0.156**	**0.111**	**0.408**	0.183	**0.360**

Table 3: Performance of TARG and related work on the DSTC9 dataset; *baseline* refers to the DSTC9 provided baseline. Numbers in **bold** denote best results in that metric.

Model	Knowledge Selection		Response Generation
	EM	F1	BLEU-4
Baseline	37.2	52.9	17.7
JARS	42.1	57.8	-
CAiRE	45.7	60.1	22.3
RWTH	46.6	62.8	24.4
TARG	**49.8**	**66.4**	**28.6**

Table 4: Performance of TARG and related work on the Doc2Dial dataset; *baseline* refers to the Doc2Dial provided baseline. Numbers in **bold** denote best results in that metric.

using a BERT (Devlin et al., 2019) encoder to predict the grounding span in the document and a BART model to generate system responses.

JARS (Khosla et al., 2021): A transformer-based (Lan et al., 2019) extractive question-answering model that extracts relevant spans from the documents. They focus on knowledge selection and do not perform response generation.

CAiRE (Xu et al., 2021a): An ensemble approach of fine-tuned RoBERTa (Liu et al., 2019) models, trained with a meta-learning objective over data-augmented datasets.

RWTH (Daheim et al., 2021): They use a biaffine classifier to model spans, followed by an ensemble for knowledge selection, and a cascaded model that grounds the response prediction on the predicted span for response generation.

DSTC9-baseline (Gunasekara et al., 2020): The baseline provided by the DSTC9 challenge is a response generation model obtained by fine-tuning the GPT-2 (Budzianowski and Vulić, 2019) model with a standard language modeling objective.

KDEAK (Chaudhary et al., 2021): A model which formulates knowledge selection as a factorized retrieval problem with three modules performing domain, entity and knowledge level analyses. The response is generated using a GPT-2 model attending on any relevant retrieved knowledge.

RADGE (Tang et al., 2021): A multi-task method that exploits correlations between dialogue history and keywords extracted from the API through fine-tuning a sequence of ELECTRA models (Clark et al., 2020).

EGR (Bae et al., 2021): An approach that uses relevance similarity to score factoids, and later reranks them with a rule-based algorithm based on entity names parsed from the dialogue. The response is generated with a BART model.

4.3 Evaluation Measures

We make use of the following automatic evaluation metrics in our experiments. For each dataset, we calculate the metrics used by the respective challenges for consistency.

Exact Match (EM): This measures what part of the predicted knowledge span matches the ground truth factoid exactly.

Token-Level F1: We cast the predicted spans and ground truth factoids as bags of tokens, and compute F1 between them.

MRR@5: A metric based on the rank of the first ground truth factoid in a system's top-5 ranking.

Recall@5: This metric counts how many ground truth factoids occur in a system's top-5 ranking.

BLEU-X (Papineni et al., 2002): BLEU-X estimates a generated response's via measuring its n-gram precision against the ground truth. X denotes the maximum size of the considered n-grams (i.e. unigrams, bigrams, trigrams, 4-grams).

ROUGE-X (Lin, 2004): ROUGE-X measures n-gram recall between generated and ground truth response. ROUGE-L measures the longest common word subsequence.

4.4 Implementation Details

We use a pre-trained BART-base model to encode utterances and factoids. The max sentence length is set to 50 and the max number of dialogue turns is set to 15. The hidden size of attentions are all

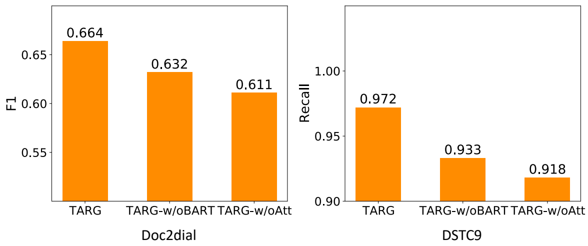

Figure 3: Ablation study for knowledge selection.

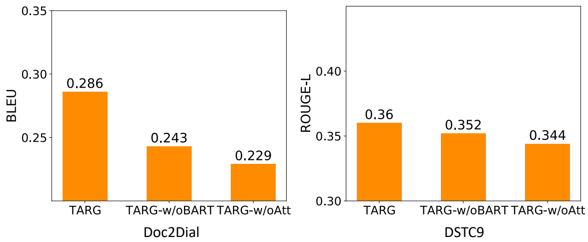

Figure 4: Ablation study for response generation.

set to 768. The size of the convolution and pooling kernels are set to (3, 3, 3). The joint loss λ is 0.5. The dropout probability is 0.1. The batch size is set to 8. We optimize with Adam and an initial learning rate of 3e-5.

4.5 Experimental Results

Table 3 and Tables 4 show our results on DSTC9 and Doc2Dial respectively. Observe that TARG performs significantly better than related work in both knowledge selection and response generation on the Doc2dial dataset, outperforming the second best system by 3.2, 3.6, and 4.2 points in EM, F1 and BLEU-4 respectively.

On the DSTC9 dataset, TARG outperforms the related work in most metrics, though by narrow margins. Due to the smaller differences, we consider TARG to be performing on par with state-of-the-art on DSTC9. The performance gains of TARG can be explained by the topic-aware mechanism as it provides a more flexible way to weight and aggregate different dialogue history turns. This indicates that better understanding of the dialogue history is crucial for predicting the relevant factoids and generating a reasonable response.

5 Discussion

5.1 Ablation Study

Here we conduct an ablation study of TARG, to explore the effects of the BART model, topic-aware attention, as well as the different topic attention mechanisms. The results indicate that all these mechanisms are necessary to the performance of knowledge selection and response generation.

Model	Knowledge Selection EM	Knowledge Selection F1	Response Generation BLEU
TARG-dot	0.468	0.642	0.261
TARG-bilinear	0.481	0.652	0.268
TARG-outer	0.489	0.655	0.275
TARG	**0.498**	**0.664**	**0.286**

Table 5: Ablation study over different attention mechanisms on knowledge selection and response generation.

Effect of BART: To investigate the effectiveness of using BART in the utterance / factoid encoder and system response generator, we replace BART with a bi-directional LSTM and rerun the model for Doc2dial and DSTC9. As shown in Figures 3 and 4, the performance of the BiLSTM-based model TARG-w/oBART decreases significantly in knowledge selection, and especially in response generation as is indicated by the drop in BLEU. As expected, this indicates that the BART model can create and utilize more accurate representations for dialogue history and unstructured knowledge.

Effect of topic-aware attention: Next we remove the topic-aware attention mechanisms (TARG-w/oAtt). Figures 3 and 4 again show that the respective performances deteriorate considerably. This shows that topic-aware attention helps derive an important weighting scheme over the utterances leading to better understanding of dialogue history.

Effect of topic attention mechanisms: Here we compare TARG against TARG-dot, TARG-bilinear, and TARG-outer which use exclusively doc product attention, bilinear attention, and outer product attention respectively. Table 5 shows that doc product attention underperforms compared to bilinear and outer product attention while bilinear attention's performance is comparable with outer product attention. In addition, any isolated attention mechanism performs considerably worse than their fusion, supporting its utilization. We conjecture that this is due to how different attention mechanisms focus on different topic features.

5.2 Case Study

Consult Figure 5 for a case study from the Doc2Dial dataset. On the top of the Figure are the previous turns of dialogue history, while on the right is a subset of the available factoids. We can observe how the topic changes throughout the turns of dialogue history (by consulting the corresponding factoid topic), from "Exploring Your

Dialogue History Turns	
U1	U: I wanted to know about career options.
S1	S: Do you love working with animals?
U2	U: No, what else you got?
S2	S: Do you like working with computers?
U3	U: I use them but wouldn't care to work on computer related things. Do you have any info for the parents to look at?
S3	S: Is this information for a parent that is planning ahead for a child's higher education?
U4	U: yes it is.
S4	S: We have resources for parents to learn more about saving early, and finding tax breaks.
U5	U: Do you have any info on how college can help me?
Generated Response	
Ground Truth	Yes, you can look at our Reasons to Attend a College or Career School section.
TARG	Please look at Reasons to Attend a College or Career School.
RWTH	Yes, Budgeting Resources for Parents of Students.
Doc2Dial-baseline	Review our resources for parents.

Knowledge Candidates (Factoids)	Topic	Context
K1	Exploring Your Career Options	Love working with animals? How about computers? Find possible careers to match your interests.
K2	Resources for Parents of Students	Are you a parent planning ahead for your child s higher education? Review our resources for parents to learn more about saving early, and finding tax breaks.
K3	Preparing for College	Check out Reasons to Attend a College or Career School. Learning About Budgeting Resources for Parents of Students.

Figure 5: Case study on Doc2Dial. Dialogue history turns are grounded to knowledge candidates of the same color.

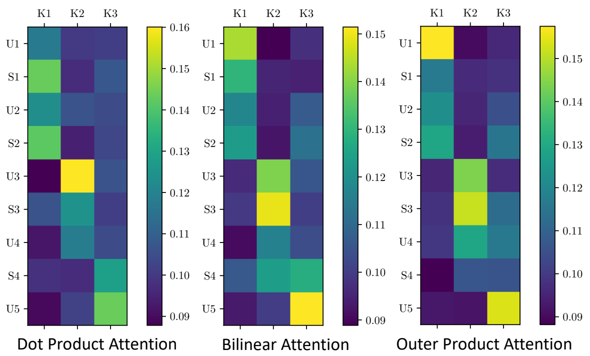

Figure 6: Visualization of learned topic-aware attention of dialogue history utterances U-X and S-X (for user and system utterance) for each factoid K-X in the example in Figure 5. Lighter spots mean higher attention scores.

Career Options" in turns 1 and 2, to “Resources for Parents of Students" in turns 3 and 4, and finally “Preparing for College" in turn 5.

On the bottom of Figure 5, we present responses generated by our proposed model TARG, the best of the previous work RWTH, the Doc2Dial-baseline, and the ground truth. Observing the responses and comparing with the ground truth, RWTH and the Doc2Dial-baselines seem to generate irrelevant responses, picking the wrong factoids from the candidates on the right, i.e. “Budgeting Resources for Parents of Students" and “Review our resources for parents". TARG generates the more relevant and fluent response of the three, as its topic-aware attention informs knowledge selection to pick the factoid that more naturally follows the dialogue history, i.e. “Reasons to Attend a College or Career School”. Furthermore, TARG’s BART decoder ensures the fluency of the output.

Figure 6 presents a visualization of TARG’s learned topic-aware attention over the dialogue utterances and factoids of the case study. This includes Dot Product Attention, Bilinear Attention, and Outer Product Attention. We can see that topic-aware attention captures reasonable dialogue utterance weights for each factoid, with the weighing moving from topic K1 to K2 and to K3 as attentions are calculated over the dialogue history utterances. This supports our claim that modeling the topic shifts can be helpful for knowledge selection, and consequently response generation, through better understanding of the dialogue history.

6 Conclusion

In this paper, we proposed TARG: “Topic-Aware Response Generation”, a topic-aware model which incorporates multiple topic-aware attention mechanisms to derive the importance weighting scheme over both dialogue utterances and unstructured external knowledge, and through that facilitate better dialogue history understanding. Our proposed method achieves state-of-the-art results in both knowledge selection and response generation, outperforming previous state-of-the-art by 3.2, 3.6, and 4.2 points in EM, F1 and BLEU-4 respectively on Doc2Dial, and performing comparably with previous work on DSTC9. To provide further insights, we also presented an ablation study of our model that supported the importance of our method’s various components, and discussed a case study accompanied by an analysis of the attention mechanisms.

References

Hyunkyung Bae, Minwoo Lee, AhHyeon Kim, Cheong-jae Lee Hwanhee Lee, Cheoneum Park, Donghyeon Kim, and Kyomin Jung. 2021. Relevance similarity scorer and entity guided reranking for knowledge grounded dialog system. *In AAAI*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. *In ICLR*.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2-how can i help you? towards the use of pre-trained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22.

Mudit Chaudhary, Borislav Dzodzo, Sida Huang, Chun Hei Lo, Mingzhi Lyu, Lun Yiu Nie, Jinbo Xing, Tianhua Zhang, Xiaoying Zhang, Jingyan Zhou, et al. 2021. Unstructured knowledge access in task-oriented dialog modeling using language inference, knowledge retrieval and knowledge-integrative response generation. *In AAAI*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *In ICLR*.

Nico Daheim, David Thulke, Christian Dugast, and Hermann Ney. 2021. Cascaded span extraction and response generation for document-grounded dialog. *In ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *In ICLR*.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. Doc2dial: A goal-oriented document-grounded dialogue dataset. In *In EMNLP*, pages 8118–8128.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *In AAAI*, volume 32.

Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zhu, and Quan Liu. 2019. Dually interactive matching network for personalized response selection in retrieval-based chatbots. In *In EMNLP*, pages 1845–1854.

Jia-Chen Gu, Zhenhua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2020. Filtering before iteratively referring for knowledge-grounded response selection in retrieval-based chatbots. In *In EMNLP*, pages 1412–1422.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. 2020. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486*.

Sopan Khosla, Justin Lovelace, Ritam Dutt, and Adithya Pratapa. 2021. Team jars: Dialdoc subtask 1-improved knowledge identification with supervised out-of-domain pretraining. In *In ACL*.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access. In *In SIGDIAL*, pages 278–289.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *In ICLR*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *In ACL*, pages 7871–7880.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *In NIPS*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, Barcelona, Spain. Association for Computational Linguistics.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *In ACL*, pages 1489–1498.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *In CoRR*.

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

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pre-trained auto-regressive model.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, William B Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *In ACL*, pages 5427–5436.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *In AAAI*, volume 34, pages 8689–8696.

Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In *In AAAI*, volume 34, pages 8697–8704.

Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, and Rui Yan. 2018. An ensemble of retrieval-based and generation-based human-computer conversation systems. In *In IJCAI*.

Yajing Sun, Yue Hu, Luxi Xing, Jing Yu, and Yuqiang Xie. 2020. History-adaption knowledge incorporation mechanism for multi-turn dialogue system. In *In AAAI*, volume 34, pages 8944–8951.

Liang Tang, Qinghua Shang, Kaokao Lv, Zixi Fu, Shijiang Zhang, Chuanming Huang, , and Zhuo Zhang. 2021. RADGE: Relevance learning and generation evaluating method for task-oriented conversational system-anonymous version. *In AAAI*.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Improving knowledge-aware dialogue generation via knowledge base question answering. In *In AAAI*, volume 34, pages 9169–9176.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. 2021. A controllable model of grounded response generation. In *In AAAI*, volume 35, pages 14085–14093.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3351–3357. AAAI Press.

Yan Xu, Etsuko Ishii, Genta Indra Winata, Zhaojiang Lin, Andrea Madotto, Zihan Liu, Peng Xu, and Pascale Fung. 2021a. Caire in dialdoc21: Data augmentation for information-seeking dialogue system. In *In ACL*.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021b. Topic-aware multi-turn dialogue modeling. In *In AAAI*.

Rui Yan and Dongyan Zhao. 2018. Coupled context modeling for deep chit-chat: towards conversations between human and computer. In *In SIGKDD*, pages 2574–2583.

Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tur. 2019. Deepcopy: Grounded response generation with hierarchical pointer networks. In *In SIGDIAL*, pages 122–132.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *In AAAI*.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018. Reinforcing coherence for sequence to sequence model in dialogue generation. In *In IJCAI*, pages 4567–4573.

Hainan Zhang, Yanyan Lan, Liang Pang, Hongshen Chen, Zhuoye Ding, and Dawei Yin. 2020. Modeling topical relevance for multi-turn dialogue generation. *In IJACI*.

Xueliang Zhao, Chongyang Tao, Wei Wu, Can Xu, Dongyan Zhao, and Rui Yan. 2019. A document-grounded matching network for response selection in retrieval-based chatbots. *In IJCAI*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *In EMNLP*, pages 3377–3390.