Q-TOD: A Query-driven Task-oriented Dialogue System

Xin Tian* **Yingzhan Lin*** **Mengfei Song*** **Siqi Bao**
Fan Wang **Huang He** **Shuqi Sun** **Hua Wu**
Baidu Inc., China
{tianxin06, linyingzhan01, songmengfei01}@baidu.com

Abstract

Existing pipelined task-oriented dialogue systems usually have difficulties adapting to unseen domains, whereas end-to-end systems are plagued by large-scale knowledge bases in practice. In this paper, we introduce a novel query-driven task-oriented dialogue system, namely Q-TOD. The essential information from the dialogue context is extracted into a query, which is further employed to retrieve relevant knowledge records for response generation. Firstly, as the query is in the form of natural language and not confined to the schema of the knowledge base, the issue of *domain adaption* is alleviated remarkably in Q-TOD. Secondly, as the query enables the decoupling of knowledge retrieval from the generation, Q-TOD gets rid of the issue of *knowledge base scalability*. To evaluate the effectiveness of the proposed Q-TOD, we collect query annotations for three publicly available task-oriented dialogue datasets. Comprehensive experiments verify that Q-TOD outperforms strong baselines and establishes a new state-of-the-art performance on these datasets.

1 Introduction

Task-oriented dialogue systems are designed to help users achieve their goals, such as restaurant reservation, calendar scheduling, and movie recommendation. Typically, these systems need to rely on external knowledge bases to retrieve necessary information for response generation (Eric et al., 2017; Wen et al., 2017; Eric et al., 2020). Some end-to-end trainable approaches try to encode the knowledge base into a memory module and attend relevant knowledge records for response generation (Wu et al., 2019; Qin et al., 2020; Raghu et al., 2021). Since these end-to-end approaches need to continually refresh the memory module, a large-scale knowledge base will lead to a heavy computation burden and difficult joint optimization. Recently, some works leverage the power of pre-trained language models and take the entire linearized knowledge base as the input to assist response generation (Gou et al., 2021; Xie et al., 2022). However, the input sequence could easily become too long to feed into the transformer network. Considering there are thousands or millions of records in industrial knowledge bases, the *knowledge base scalability* becomes a critical challenge for these end-to-end approaches.

In these circumstances, the practical deployed systems tend to employ pipelined designs and strip out the component of knowledge retrieval (Wen et al., 2017; Hosseini-Asl et al., 2020; Su et al., 2022). These pipelined systems usually consist of natural language understanding, dialogue state tracking, dialogue policy learning, and system response generation. In order to retrieve relevant information from the external knowledge base, these approaches need to pre-define the schema of dialogue states according to the knowledge base. Due to this kind of strong association, these pipelined systems have difficulties adapting to unseen domains, i.e., weak ability on *domain adaptation*.

To tackle these issues, in this paper, we introduce a novel Query-driven Task-oriented Dialogue (Q-TOD) system. The overview of Q-TOD is shown in Figure 1, where three sequential modules are included: 1) the *query generator* extracts the essential information from the dialogue context into a concise query in an unstructured format of the natural language; 2) the generated query is then utilized to retrieve relevant knowledge records with an off-the-shelf *knowledge retriever*; 3) the *response generator* produces a system response based on the retrieved knowledge records and the dialogue context.

The advantages brought by the query-driven task-oriented dialogue system are two-fold. Firstly, the query is in the unstructured format of natural language, which is not confined to the knowledge base and is able to mitigate the issue of *domain adapta-*

*Equal contribution.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7260 - 7271

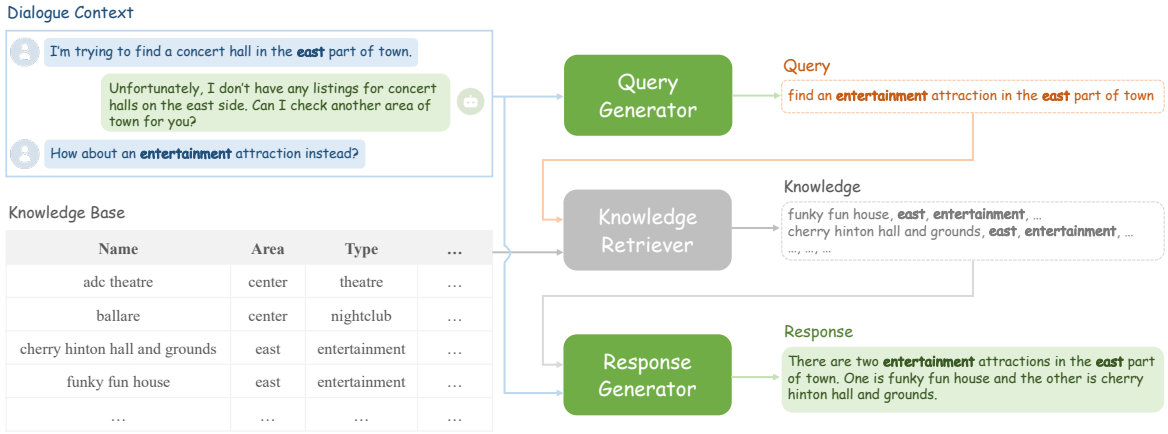

Figure 1: The overview of the proposed Q-TOD. Q-TOD consists of three modules, which are invoked sequentially: query generator, knowledge retriever, and response generator. Query generator and response generator are trained with a shared transformer, whereas knowledge retriever is an off-the-shelf retrieval model, allowing plug-and-play modularity.

tion. Secondly, with the incorporation of the query, Q-TOD decouples the knowledge retrieval from the response generation, getting rid of the issue of the *knowledge base scalability*. To explore the effectiveness of the query-driven systems, we collect query annotations for three public task-oriented dialogue datasets: SMD (Eric et al., 2017), CamRest (Wen et al., 2017), and MultiWOZ-2.1 (Eric et al., 2020). Experimental results demonstrate that Q-TOD achieves superior performance as compared to other state-of-the-art approaches. Particularly, in the few-shot settings, Q-TOD achieves a comparable performance with the previous state-of-the-art using only 5% of the training data. Our collected data, code, and models have been released at GitHub[1], hoping to facilitate further research in task-oriented dialogue systems.

2 Methodology

The goal of this paper is to explore a novel and effective framework for task-oriented dialogue systems. As shown in Figure 1, the proposed Q-TOD consists of three subsequent modules: query generator, knowledge retriever, and response generator. The detailed design of these three modules will be discussed in the following.

2.1 Query Generator

The *query generator* aims to extract essential information from the dialogue context into a natural language query. For a multi-turn conversation, the dialogue context at the t-th turn can be represented as $C_t = \{U_0, R_0, \ldots, U_t\}$, where each turn consists of user utterance U_i and system response R_i. With the dialogue context as input, the query at the t-th turn is generated with a Transformer-based language model:

$$Q_t = \mathit{Transformer}\,(C_t) \tag{1}$$

In the query generation, the noisy or out-of-date information from the context is supposed to be removed, whereas the essential and up-to-date requirements raised by the user should be highlighted. As shown in the example of Figure 1, the query only contains the minimal user requirements in the current turn (i.e., find an east entertainment attraction) and discards the outdated attraction type of concert hall. In cases where no query is required, e.g. greetings or thanks, a special token `[NOTHING]` is used to represent the null query at this turn.

Recently, in some knowledge-intensive conversations, there is a trend to employ the query to enhance the performance of relevant knowledge retrieval. In conversational question answering, to deal with ellipsis and coreference, a question rewriting task is introduced to convert a context-dependent question into a self-contained query (Vakulenko et al., 2021; Anantha et al., 2021). In open-domain knowledge-grounded dialogue, to incorporate real-time external information, some works learn to generate a search query and leverage search engines for response generation (Komeili et al., 2022; Shuster et al., 2022). To the best of our knowledge, Q-TOD is the first work that tries to encode the natural language query into a task-

[1] https://github.com/PaddlePaddle/Knover/tree/develop/projects/Q-TOD

oriented dialogue system. Distinct from the above approaches, the query in Q-TOD is designed to extract the essential and up-to-date user requirements.

2.2 Knowledge Retriever

The *knowledge retriever* utilizes the generated query to retrieve relevant knowledge records from the external knowledge base:

$$K_t = \textit{Retriever}\,(Q_t; \mathcal{K}) \tag{2}$$

where $\mathcal{K}$ refers to the entire knowledge base, and $K_t = \{k_t^1, k_t^2, \ldots, k_t^n\}$ are retrieved top-n relevant knowledge records. As displayed in Figure 1, the module of knowledge retriever is a black box in this system. In fact, any off-the-shelf knowledge retriever can be employed in Q-TOD, including BM-25 or dense retrieval models. Such strong adaptability and flexibility mainly benefit from the preceding query generation. Firstly, the query is in the format of natural language, which is a universal representation and adaptable to commonly used retrievers. Secondly, although multi-turn dialogue context typically requires elaborately designed or tuned retrievers (Shuster et al., 2021), by extracting essential information into a concise query, the off-the-shelf retriever can achieve relatively good performance as well.

In fact, the module of knowledge retriever is the key to the knowledge base scalability of Q-TOD. Given a large-scale knowledge base, the retriever filters out massive irrelevant knowledge records and picks out top-n relevant ones. In this way, the subsequent response generation is not affected by the size of the knowledge base and is able to pay more attention to knowledge utilization. As suggested by recent works (Adolphs et al., 2021; Shuster et al., 2022) and verified in our experiments, the decoupling of knowledge retrieval and response generation alleviates the modeling difficulty and boosts the final performance.

2.3 Response Generator

The *response generator* produces the system response given the retrieved top-n knowledge records and dialogue context:

$$R_t = \textit{Transformer}\,(K_t; C_t) \tag{3}$$

With the assistance of the preceding modules, the response generator can focus more on precise knowledge utilization and produce high-quality replies towards the dialogue context.

Statistics	SMD	CamRest	MWOZ
Dialogues	3031	676	2097
Utterances	15928	5488	19632
Domains	3	1	3
Turns per Dialogue	5.26	8.12	8.89
Tokens per Utterance	7.97	12.31	14.73
Tokens per Query	7.50	9.67	10.57

Table 1: Statistics of the datasets in the experiments.

In Q-TOD, we train the query generator and response generator jointly with a shared transformer. To distinguish these tasks in a single model, two task-specific discrete prompts Z_Q and Z_R are adopted and concatenated with the rest input. For query generation, the prompt Z_Q is "`translate dialogue context to query:`". For response generation, the prompt Z_R is "`generate system response based on knowledge and dialogue context:`". Overall, the training objective is to minimize the following negative log-likelihood loss:

$$\begin{aligned}\mathcal{L} = &-\log P_\Theta\,(Q_t|Z_Q; C_t) \\ &-\log P_\Theta\,(R_t|Z_R; K_t; C_t)\end{aligned} \tag{4}$$

The knowledge retrieval is a black box in this system and thus not involved in the optimization.

3 Experiments

3.1 Datasets

To investigate the effectiveness of the proposed query-driven dialogue system, we collect query annotations for three publicly available multi-turn task-oriented dialogue datasets: Stanford Multi-Domain (SMD) (Eric et al., 2017), CamRest (Wen et al., 2017), and MultiWOZ-2.1 (MWOZ) [2] (Eric et al., 2020). The query annotations are collected through three stages. The authors first provide ten examples of dialogue sessions with query annotations for each dataset. Then, the crowd workers complete all query annotations on these dialogues after reading the examples. Finally, to ensure the quality of annotations, multiple data specialists will review it. We will release the collected data for further research.

[2]For MultiWOZ-2.1, following previous works (Qin et al., 2020; Raghu et al., 2021), we use the version released by Qin et al. (2020), which equips each dialogue with corresponding knowledge base.

Model	SMD		CamRest		MWOZ	
	Entity F1	BLEU	Entity F1	BLEU	Entity F1	BLEU
DSR	51.90†	12.70†	53.60†	18.30†	30.00‡	9.10‡
KB-Retriever	53.70	13.90	58.60	18.50	-	-
GLMP	60.70‡	13.90‡	58.90§	15.10§	32.40‡	6.90‡
DF-Net	62.70	14.40	-	-	35.10	9.40
GPT-2+KE	59.78	17.35	54.85	18.00	39.58	15.05
CDNET	62.90	17.80	68.60	21.80	38.70	11.90
COMET	63.60	17.30	-	-	-	-
UnifiedSKG (T5-Large)	65.85	17.27	71.03*	20.31*	46.04*	13.69*
UnifiedSKG (T5-3B)	67.88	15.45	72.78*	18.46*	49.65*	13.01*
Q-TOD (T5-Large)	71.11	21.33	74.22	23.75	50.61	17.62
Q-TOD (T5-3B)	**73.44**	**21.76**	**76.81**	**24.65**	**53.28**	**18.27**

Table 2: Experimental results on the SMD, CamRest, and MWOZ datasets, with the best value written in bold [3]. †, ‡, § denotes that the results are cited from Qin et al. (2019), Qin et al. (2020), and Raghu et al. (2021), respectively. $*$ indicates that we reproduce the results using the official code released by the authors.

In our experiments, we utilize the provided training/validation/test partitions of all benchmark datasets. Table 1 summarizes the statistics of the above three datasets.

3.2 Experimental Settings

Our experiments are carried out with T5 (Raffel et al., 2020), in which two model sizes are used: T5-Large and T5-3B. For knowledge retriever, we leverage an off-the-shelf retrieval model RocketQA (Ren et al., 2021). Particularly, we fine-tune T5 with AdamW optimizer (Loshchilov and Hutter, 2019) and Noam learning rate scheduler (Vaswani et al., 2017). During inference, the decoding strategy of beam search is employed, with a beam size of 4. And the number of retrieved knowledge records top-n is set to 3. All the models are trained on 8 NVIDIA Tesla A100 GPU cards for 50 epochs and early stopped according to the performance on the validation set. More details about hyper parameter settings are provided in Appendix A.

3.3 Baselines

We compare Q-TOD with the following strong baselines:

DSR (Wen et al., 2018) models dialogue state as distributed representation to query the knowledge base with an attention mechanism.

KB-Retriever (Qin et al., 2019) proposes an entity-consistency augmented decoder to focus on a single row of the knowledge base by memory network and attention mechanism.

GLMP (Wu et al., 2019) leverages a global-to-local pointer network to first generate a sketch response and then fill slots with entities from the knowledge base.

DF-Net (Qin et al., 2020) applies the Mixture-of-Experts mechanism (MoE) to dynamically exploit the relevance between the target domain and all source domains.

GPT-2+KE (Madotto et al., 2020) proposes to pack the knowledge base into the model parameters implicitly through dialogue data augmentation.

CDNET (Raghu et al., 2021) computes a distillation distribution over the knowledge records, which is used to get the final copy distribution for entity choosing.

COMET (Gou et al., 2021) introduces a Memory-Masked Encoder to enforce entities interact within the same knowledge record, aiming to avoid the distraction from the irrelevant ones.

UnifiedSKG (Xie et al., 2022) recasts 21 structured knowledge grounding tasks into a unified text-to-text language model (including task-oriented dialogue modeling) and achieves state-of-the-art performance on these tasks.

3.4 Results

Following the prior works (Eric et al., 2017; Wu et al., 2019; Qin et al., 2020; Madotto et al., 2020; Raghu et al., 2021; Gou et al., 2021; Xie et al., 2022), we evaluate the model performance with metrics of micro **Entity-F1** (Eric et al., 2017) and

[3] Since UnifiedSKG includes a base size version on SMD, we also train a T5-Base Q-TOD for comparison. Q-TOD (T5-Base) obtains 68.22% Entity-F1 and 20.14% BLEU, which outperforms UnifiedSKG on the same scale (UnifiedSKG T5-Base, 66.45% Entity-F1, 17.41% BLEU).

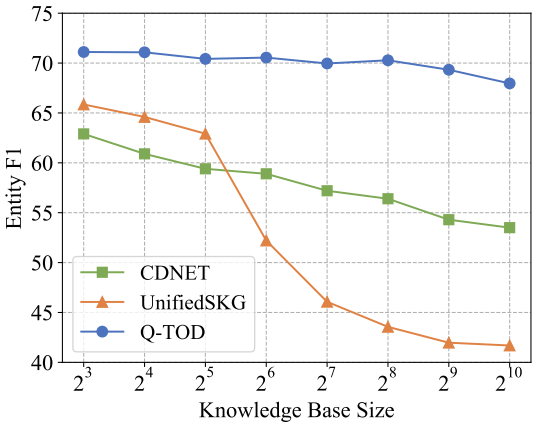

Figure 2: Effects of knowledge base scaling.

BLEU (Papineni et al., 2002). The Entity-F1 measures the model's abilities to generate relevant entities from the external knowledge base according to the dialogue context. BLEU measures the n-gram overlap between generated response and the oracle response.

Table 2 summarizes the results of Q-TOD and all baselines on three datasets. It can be observed that our Q-TOD consistently outperforms all previous models, achieving a new state-of-the-art result. Specifically, on the Entity-F1 metric, Q-TOD achieves the absolute improvement of 5.56% on SMD, 4.03% on CamRest, and 3.36% on MWOZ, respectively. On the BLEU metric, Q-TOD also obtains the highest score with the increment of 3.96%, 2.85%, and 3.22% on SMD, CamRest, and MWOZ, respectively [4]. These results demonstrate that the proposed Q-TOD can generate high-quality system responses with the relevant knowledge records. It is worth noting that, UnifiedSKG (Xie et al., 2022) employs a Transformer-based response generator similar to ours, but our model surpasses it on all three datasets under both T5-Large and T5-3B sizes. This confirms the benefits of our proposed query-driven retrieval, which helps the response generator avoid distractions from irrelevant knowledge records and focus on the utilization of relevant ones.

4 Discussion

For further analysis of Q-TOD, we will have discussions on the following aspects: knowledge base scalability, effect of query, performance of precise knowledge, domain adaptation, and case study. In this section, unless specified, experiments are carried out with T5-Large.

4.1 Knowledge Base Scalability

In practice, the knowledge base of a specific domain usually contains thousands of records, such as weather forecasts and music recommendations. Hence, it is necessary to explore the knowledge base scalability in task-oriented dialogue systems. To this end, we simulate large knowledge bases by expanding the existing ones. Specifically, we expand the original session-level knowledge bases on the SMD dataset by injecting dataset-level knowledge records and some crawled knowledge records [5].

As shown in Figure 2, we compare Q-TOD with two strong baselines, CDNET (Raghu et al., 2021) and UnifiedSKG (Xie et al., 2022). The results show that when increasing the knowledge base size from 2^3 to 2^{10}, Q-TOD is able to maintain a stable performance on Entity F1. In particular, when the size of the knowledge base is expanded to 2^{10} (128 times compared with the original SMD), Q-TOD obtains 67.96% Entity F1, only a decrease of 3.15%. The superior performance is owing to the fact that Q-TOD extracts the essential information from the dialogue context into the query. The short query enables the knowledge retrieval to be decoupled from the response generation, getting rid of the issue of the *knowledge base scalability*. In contrast, the performance of CDNET and UnifiedSKG decreases gradually as the size of the knowledge base increases. This indicates that joint modeling can barely adapt to large-scale knowledge bases, which might result from the difficulty of a single model in handling implicit knowledge retrieval and response generation simultaneously. Moreover, due to the limitation of the max input length in UnifiedSKG, the Entity F1 decreases sharply when the size of the knowledge base is greater than 2^6.

4.2 Effect of Query

To investigate the effectiveness of query incorporation, we would like to answer the following two research questions (RQ).

RQ1: *what would happen to Q-TOD without query annotations?*

[4]The three datasets we used include only one reference system response, which results in comparably low BLEU scores for Q-TOD and all baselines.

[5]The average size of the session-level knowledge base for SMD is 7.19. We construct a dataset-level knowledge base by merging all session-level knowledge bases, where the average size of the dataset-level knowledge base is 542.67. In our experiments, extra crawled knowledge records are added to the knowledge base when necessary.

Model	Entity F1
UnifiedSKG	65.85
Q^I-TOD	69.57
Q-TOD	71.11

Table 3: Comparison of performance between UnifiedSKG, Q^I-TOD, and Q-TOD on SMD.

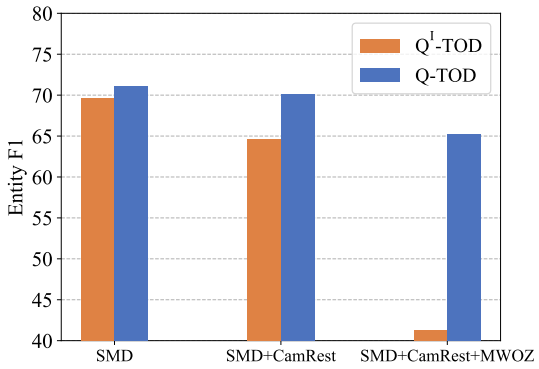

Figure 3: Comparison of performance between Q^I-TOD and Q-TOD.

Under our framework, to deal with the situation without any query annotation, a straightforward solution is to degrade the query generator to identical mapping. In other words, the dialogue context itself can be regarded as a naive query to retrieve knowledge records for response generation. This setting with identical mapping is denoted as Q^I-TOD. Experiments with Q^I-TOD are carried out on SMD and results are summarized in Table 3. Q^I-TOD obtains 69.57% Entity F1, outperforming the previous state-of-the-art UnifiedSKG (65.85% Entity F1). This indicates that our framework can also achieve state-of-the-art even without query annotations.

RQ2: *why not content with Q^I-TOD?*

In practical deployments, the conversations are more complex and noisy as compared to those in public datasets. For instance, one user might ask for navigation after a restaurant reservation, also known as a cross-domain conversation. Since the context contains distractions of the noisy or outdated information, Q^I-TOD encounters difficulties in retrieving relevant knowledge records. For better comparison between Q^I-TOD and Q-TOD, we further construct two cross-domain dialogue datasets by merging dialogue sessions from SMD, CamRest, and MWOZ. Detailed construction process is described in Appendix D. As shown in Figure 3, the gap between Q^I-TOD and Q-TOD becomes larger on cross-domain datasets, especially more than 20% Entity-F1 on `SMD+CamRest+MWOZ`.

Model	Entity F1
Q-TOD (T5-Large)	71.11
w/ fine-tuned retriever	71.17 (+0.06)
w/ oracle knowledge	71.96 (**+0.85**)
Q-TOD (T5-3B)	73.44
w/ fine-tuned retriever	74.96 (+1.52)
w/ oracle knowledge	76.20 (**+2.76**)

Table 4: Performance of more precise knowledge. *Oracle knowledge* refers to taking the ground-truth of knowledge as the input of the response generator, which is the theoretical upper bound.

The limited performance of knowledge retrieval of Q^I-TOD results in low-quality system responses in cross-domain scenarios. These results suggest that the query generator is a crucial module in our framework, which ensures that the noisy or out-of-date information is filtered out and will not be transmitted to the knowledge retriever.

4.3 Performance of Precise Knowledge

Although the off-the-shelf knowledge retriever is utilized in Q-TOD, a domain-specific knowledge retriever can be employed to obtain precise knowledge when the knowledge annotation is available. To investigate the effect of precise knowledge, we collect turn-level knowledge annotations to fine-tune a new knowledge retriever on the SMD dataset. Here, the knowledge retriever is initialized with T5 (Raffel et al., 2020), where the model takes the query and each linearized knowledge record as input and outputs a relevance label: `MATCHED` or `MISMATCHED`.

The results are summarized in Table 4. It can be observed that the system with fine-tuned knowledge retriever achieves better Entity-F1 than the off-the-shelf retriever. To give an idea of the performance limits of knowledge retrieval, we also evaluate it with oracle knowledge records. The performance with oracle knowledge shows that there is some headroom for Q-TOD when improving the knowledge precision. Moreover, we note that compared with T5-Large, T5-3B achieves more improvements using more precise knowledge, especially the improvement of Entity F1 becomes 2.76% with oracle knowledge. From the insights of previous works (Thoppilan et al., 2022), this might thank to a greater ability of knowledge utilization in a larger model.

Model	Zero-Shot		Few-Shot 1%		Few-Shot 5%		Few-Shot 20%	
	Entity F1	BLEU	Entity F1	BLEU	Entity F1	BLEU	Entity F1	BLEU
PPTOD†	10.24	0.73	19.98	6.19	23.93	7.92	29.64	9.79
UnifiedSKG†	28.62	7.61	36.66	10.70	41.57	12.72	45.37	13.04
Q-TOD	**31.52**	**10.14**	**39.19**	**11.35**	**45.99**	**14.27**	**48.54**	**16.51**

Table 5: Performance of zero-shot and few-shot settings. The average scores over three runs are reported. † denotes that we reproduce the results using the official code released by the authors.

Dialogue Context	U: I'm looking for an **expensive** eastern european place in the **south**. S: Unfortunately, I don't happen to have any listing that meets what you were looking for. Would you like me to try either a different area or type of restaurant? U: How about a **chinese** restaurant?
Generated Query	find an **expensive chinese** restaurant in the **south** part of the city
Retrieved Knowledge	peking restaurant, **chinese**, **south**, **expensive**, ... the good luck chinese food takeaway, **chinese**, **south**, **expensive**, ... ugly duckling, **chinese**, centre, **expensive**, ...
Generated Response	There are two expensive chinese restaurants in the south part of town. Peking restaurant and the good luck chinese food takeaway.
Oracle Response	Certainly, we have two expensive chinese restaurants the good luck chinese food takeaway, and peking restaurant. Would you like a reservation at one, if so for how many?

Table 6: Cherry-picked case of Q-TOD. Blue indicates the essential and updated information in the dialogue context.

4.4 Domain Adaption

A primary advantage of our framework is its strong ability on domain adaptation. To this end, we compare Q-TOD with two strong baselines in zero/few-shot settings, PPTOD (Su et al., 2022) and UnifiedSKG (Xie et al., 2022). PPTOD achieves the state-of-the-art performance in pipelined dialogue systems, which relies on the predefined schema to train a plug-and-play model. UnifiedSKG is the previous state-of-the-art model in end-to-end dialogue systems, which takes the entire knowledge base as input for response generation.

4.4.1 Zero-Shot Setting

To investigate the performance of Q-TOD in zero-shot settings, we train the model on SMD and CamRest, and then evaluate the performance on the MWOZ test set. As seen in Table 5, Q-TOD significantly surpasses the two baselines on both Entity F1 and BLEU metrics. For PPTOD, since the method depends on the predefined schema, the performance is poor on unseen domains. Without any training data, Q-TOD exceeds the baseline DSR (Wen et al., 2018) in the full-training setting (31.52% vs. 30.00% on Entity F1). These results verify the strong ability on domain adaptation of the proposed framework.

4.4.2 Few-Shot Setting

To further explore the performance of Q-TOD with a small number of training samples, we evaluate it and baselines in few-shot settings. Specifically, the following three steps are carried out: training on SMD and CamRest, training on partial MWOZ training set, and evaluating on MWOZ test set. As shown in Table 5, our framework consistently outperforms all baseline models. This demonstrates that Q-TOD is capable of transferring knowledge from other domains and achieves better performance in low-resource scenarios. Notably, with only 5% of the training data, Q-TOD achieves a comparable performance with the previous state-of-the-art model UnifiedSKG (Xie et al., 2022) (45.99% vs. 46.06% on Entity F1).

4.5 Case Study

For case study, we select and present two examples of Q-TOD, including the retrieved knowledge records and generated system responses.

A cherry-picked case generated by Q-TOD is shown in Table 6. It can be observed that the generated query successfully extracts the essential and up-to-date user requirements. With this query, the off-the-shelf retriever outputs relevant and precise knowledge records. Then the response generator produces a high-quality system response based on

Dialogue Context	U: I am looking for a place to stay. The hotel should have a star of 2 and should be in the moderate price range.
Generated Query	find a moderately priced 2 star hotel
Retrieved Knowledge	ashley hotel, north, moderate, 2 star, ... lovell lodge, north, moderate, 2 star, ... a and b guest house, east, moderate, 4 star, ...
Generated Response	There are 2 hotels that meet your criteria. **Would you like to stay in the north or the centre?**
Oracle Response	There are two such hotels in the north area. The first is the ashley hotel and the second is the lovell lodge. Do you have a preference?

Table 7: Lemon case of Q-TOD. Red refers to the factually incorrect statements in the generated system response.

the retrieved knowledge records and the dialogue context.

Table 7 describes a lemon case of Q-TOD. Despite that the retrieved knowledge records are correct, the response generator produces a factually incorrect system response. The system provides an option to the user, which doesn't exist in fact. This suggests that Q-TOD also suffers from the well-known problem of knowledge hallucination, where it generates plausible looking responses that are factually incorrect.

5 Related Work

In task-oriented dialogue systems, there is a trend to develop end-to-end trainable approaches to incorporate the external knowledge base for response generation (Wen et al., 2018; Qin et al., 2019; Wu et al., 2019; Qin et al., 2020; Madotto et al., 2020; Raghu et al., 2021; Gou et al., 2021; Xie et al., 2022). Some works encode the entire knowledge base into a memory module and learn to attend to the relevant knowledge entities for response generation (Wen et al., 2018; Qin et al., 2019; Wu et al., 2019; Qin et al., 2020; Raghu et al., 2021). Recently, with the advances in the pre-trained language models, some works take the entire linearized knowledge base as the transformer input and generate the final system response directly (Gou et al., 2021; Xie et al., 2022). Additionally, Madotto et al. (2020) proposes to store the knowledge base in the language model parameters implicitly through dialogue augmentation. However, considering the knowledge base usually contains thousands of records in practice, these end-to-end trainable approaches face the critical challenges of knowledge base scalability.

Meanwhile, there are many pipelined task-oriented dialogue systems, which strip out the component of knowledge retrieval (Young et al., 2013; Wen et al., 2017; Hosseini-Asl et al., 2020; Lin et al., 2020; Yang et al., 2021; Sun et al., 2022; Su et al., 2022; He et al., 2022). They usually decompose a task-oriented dialogue system into several pipelined modules: natural language understanding, dialogue state tracking, dialogue policy learning, and system response generation (Young et al., 2013; Wen et al., 2017). Some recent works formulate all pipelined modules as a cascaded generation task using pre-trained language model (Hosseini-Asl et al., 2020; Lin et al., 2020; Yang et al., 2021; Peng et al., 2021; Lee, 2021; Sun et al., 2022). To further boost the performance, some models attempt to introduce the pre-training strategy into task-oriented dialogue systems (Liu et al., 2021; Su et al., 2022; He et al., 2022). However, these pipelined systems rely on the predefined schema to retrieve knowledge from an external knowledge base, leading to difficulties in adapting to unseen domains.

In other research fields, there are also some works using a query to store essential information or retrieve relevant knowledge. In knowledge-grounded open-domain dialogue, to incorporate real-time external information, recent works learn to generate a search query based on the dialogue context for internet searching (Komeili et al., 2022; Adolphs et al., 2021; Shuster et al., 2022). In open-domain conversational question answering, to handle the reference problem and optimize retrieval performance, recent systems introduce a question rewriting task to convert a context-dependent question into a self-contained question (Vakulenko et al., 2021; Anantha et al., 2021; Wu et al., 2021). In context-dependent text-to-SQL task, some works learn to reformulate multi-turn conversational questions into a self-contained question, and then a context-independent text-to-SQL parser follows (Chen et al., 2021; Xiao et al., 2022). To the best of our knowledge, Q-TOD is the first framework that introduces the query into task-oriented

dialogue systems.

6 Conclusion

In this paper, we propose a novel query-driven task-oriented dialogue system, namely Q-TOD. Q-TOD consists of three modules: the query generator extracts the essential and up-to-date information from the dialogue context into a concise query, the off-the-shelf knowledge retriever utilizes the generated query to retrieve relevant knowledge records, and the response generator produces the final system response using the retrieved knowledge records and the dialogue context. Comprehensive experiments show that Q-TOD consistently outperforms all baselines on three active task-oriented dialogue datasets, achieving a new state-of-the-art performance.

Limitations

It is known that large-scale generation models are hindered by inference inefficiency. In Q-TOD, the query generator and the response generator are invoked sequentially, which inevitably increases the inference latency. Besides, similar to the previous works (Roller et al., 2021; Shuster et al., 2021), Q-TOD also suffers from the knowledge hallucination problem. These findings suggest that further research should be undertaken to explore more efficient inference strategy and high-fidelity knowledge-grounded response generation.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful and constructive comments. We also thank Wen Huang, Shiwei Huang, and Jingzhou He for their help in resource coordination. This work was supported by the National Key Research and Development Project of China (No. 2018AAA0101900).

References

Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur D. Szlam, and Jason Weston. 2021. Reason first, then respond: Modular generation for knowledge-infused dialogue. *ArXiv*, abs/2111.05204.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Zhi Chen, Lu Chen, Hanqi Li, Ruisheng Cao, Da Ma, Mengyue Wu, and Kai Yu. 2021. Decoupled dialogue modeling and semantic parsing for multi-turn text-to-SQL. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3063–3074, Online. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Yanjie Gou, Yinjie Lei, Lingqiao Liu, Yong Dai, and Chunxu Shen. 2021. Contextualize knowledge bases with transformer for end-to-end task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4310, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. *ArXiv*, abs/2107.07566.

Yohan Lee. 2021. Improving end-to-end task-oriented dialog system with a simple auxiliary task. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1296–1303, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

Qi Liu, Lei Yu, Laura Rimell, and Phil Blunsom. 2021. Pretraining the noisy channel model for task-oriented dialogue. *Transactions of the Association for Computational Linguistics*, 9:657–674.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Learning knowledge bases with parameters for task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2372–2394, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. 2019. Entity-consistent end-to-end task-oriented dialogue system with KB retriever. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 133–142, Hong Kong, China. Association for Computational Linguistics.

Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu. 2020. Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6344–6354, Online. Association for Computational Linguistics.

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.

Dinesh Raghu, Atishya Jain, Mausam, and Sachindra Joshi. 2021. Constraint based knowledge base distillation in end-to-end task oriented dialogs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5051–5061, Online. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur D. Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. *ArXiv*, abs/2203.13224.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

Haipeng Sun, Junwei Bao, Youzheng Wu, and Xiaodong He. 2022. Bort: Back and denoising reconstruction for end-to-end task-oriented dialog. *ArXiv*, abs/2205.02471.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *ArXiv preprint*, abs/2201.08239.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 355–363.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Haoyang Wen, Yijia Liu, Wanxiang Che, Libo Qin, and Ting Liu. 2018. Sequence-to-sequence learning for task-oriented dialogue with dialogue state representation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3781–3792, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, and Gaurav Singh Tomar. 2021. Conqrr: Conversational query rewriting for retrieval with reinforcement learning. *ArXiv preprint*, abs/2112.08558.

Dongling Xiao, Linzheng Chai, Qian-Wen Zhang, Zhao Yan, Zhoujun Li, and Yunbo Cao. 2022. Cqr-sql: Conversational question reformulation enhanced context-dependent text-to-sql parsers. *arXiv preprint arXiv:2205.07686*.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *ArXiv preprint*, abs/2201.05966.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14230–14238.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

A Hyper Parameters

The settings of the hyper parameters used in our experiments are summarized in Table 8.

Parameters	SMD	CamRest	MWOZ
Optimizer	AdamW	AdamW	AdamW
LR Scheduler	Noam	Noam	Noam
LR	3e-5	1e-5	3e-5
Batch Size	128	128	128
Epoch	50	50	50
Beam Size	4	4	4
Input Length	1024	1024	1024
Output Length	128	128	128

Table 8: Hyper parameters used for SMD, CamRest, and MWOZ.

B Domain-wise Performance

For detailed analysis, we also provide the performance of Q-TOD on each domain of SMD and MWOZ in Table 9.

Domain	T5-Large	T5-3B
SMD Schedule	81.42	84.22
SMD Navigate	62.91	62.72
SMD Weather	69.18	73.17
MWOZ Hotel	45.25	46.71
MWOZ Attraction	54.81	62.68
MWOZ Restaurant	55.78	58.90

Table 9: Domain-wise Entity F1 of Q-TOD on SMD and MWOZ.

C Exploration on the Number of Retrieved Knowledge Records

In our experiments, the knowledge retriever outputs top-n relevant knowledge records for response generation. To explore the effect of adjusting the number of retrieved knowledge records, we evaluate Q-TOD in three different top-n settings on the validation set of SMD. In Table 10, it is observed that the model using top-3 retrieved knowledge records achieves the best Entity F1 on both T5-Large and T5-3B. This suggests that a small number of n might be inadequate to cover the necessary knowledge records for response generation. In contrast, a large number of n would inevitably introduce more noisy knowledge records and increase the difficulty of knowledge utilization in response generation.

Model	Top-1	Top-3	Top-5
Q-TOD (T5-Large)	68.98	**70.77**	70.36
Q-TOD (T5-3B)	70.29	**72.34**	71.91

Table 10: Performance of Q-TOD with top-n retrieved knowledge records on the validation set of SMD.

D Cross-domain Dataset Construction

We construct two cross-domain dialogue datasets by merging single-domain dialogue sessions from SMD, CamRest, and MWOZ. Table 11 describes the construction details for these two datasets. For instance, in `SMD+CamRest`, two original dialogue sessions are merged into a cross-domain session, where one is from SMD and the other is from CamRest. Especially, these sessions are concatenated in a random order. The two cross-domain datasets both contain 600 dialogue sessions, in which the partition of train/validation/test is 400/100/100.

	SMD+CamRest	SMD+CamRest+MWOZ
session A	navigate (SMD) schedule (SMD) weather (SMD)	navigate (SMD) schedule (SMD)
session B	restaurant (CamRest)	restaurant (CamRest) restaurant (MWOZ)
session C	-	weather (SMD) attraction (MWOZ) hotel (MWOZ)

Table 11: Detailed cross-domain dataset construction.