TOAD: Task-Oriented Automatic Dialogs with Diverse Response Styles

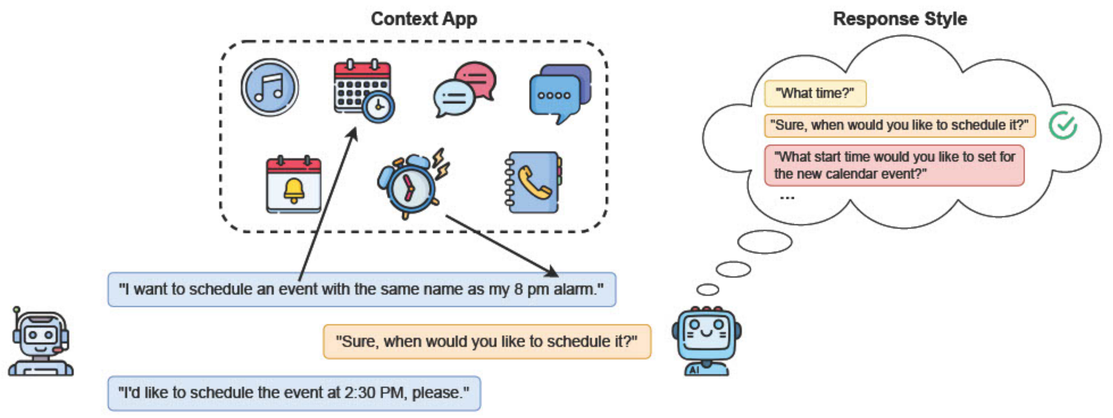
Yinhong Liu, Yimai Fang, David Vandyke, Nigel Collier
ACL 2024 · Apple and University of Cambridge

Abstract

In light of recent advances in large language models (LLMs), the expectations for the next generation of virtual assistants include enhanced **naturalness** and **adaptability** across diverse usage scenarios.

However, the creation of high-quality annotated data for TaskOriented Dialog (TOD) is recognized to be slow and costly. To address these challenges, we introduce Task-Oriented Automatic Dialogs (TOAD), **a novel and scalable TOD dataset along with its automatic generation pipeline**. The TOAD dataset simulates realistic **app context interaction** and provide a variety of system **response style** options. Two aspects of system response styles are considered, verbosity level and users' expression mirroring.

We benchmark TOAD on two response generation tasks, and the results show that modeling more verbose responses or responses without user expression mirroring is more challenging.



Methods

Data Generation Pipeline Overview

Stage 1: Persona-Grounded Context Generation

- Synthetic name, age, gender, personality.
- Synthetic occupation, marital status and hobbies.
- Synthetic personal app context instances, such as calendar, massager, reminders and music.

Stage 2: Schema-Guided Plot Generation

- Schema: TOAD supports 11 services, such as calendar, alarms, music, hotel booking and weather checking. Each service includes multiple operation intents such as create, check, delete and modify.
- Plot Representation: TOAD employs code-like meaning representations for the dialog plot. This is because 1) they can easily be automatically composed, and 2) LLM can understand the code-like syntax well.
- Context Interaction: TOAD simulates complex interactions with the app context. For example, creating a calendar event, sending message about a reminder information.

Stage 3: Dialog Generator

TOAD employs LLM to alternatively simulate user and system roles. For each system turn, TOAD generates 6 response options with distinct styles.

- Verbosity: Usually, a clear and concise communication is preferred for efficient delivery. However, in some scenario, a higher verbosity is preferred to provide comprehensive explanations and avoid ambiguity, e.g. driving or no screen is available.
- Mirroring: Copying user's expression is generally a favorable strategy for a natural conversation. However, mirroring maybe inappropriate in some cases, for example, emotional, biased or nonfactual user expressions.

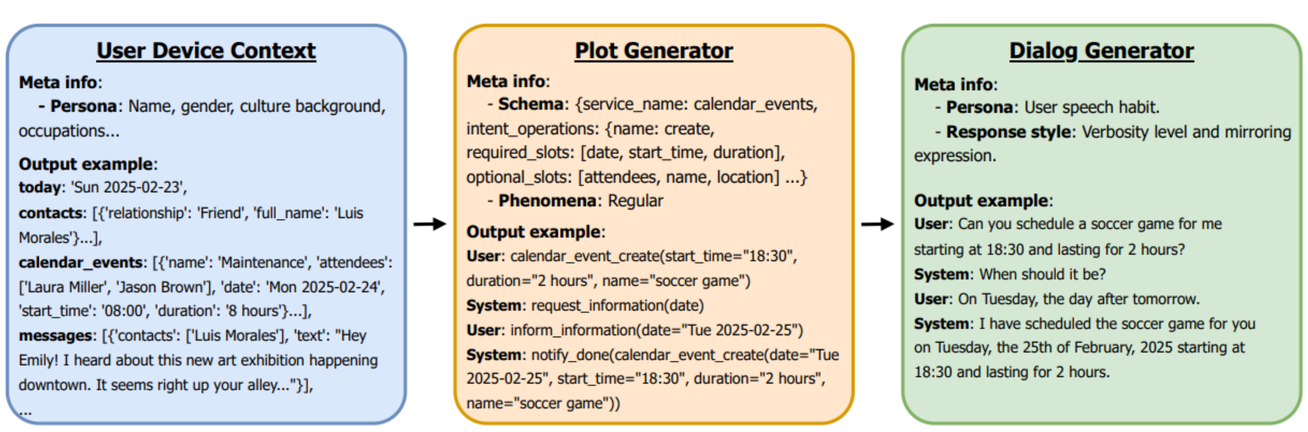


Figure 2: Overview of the TOAD Automatic Generation Pipeline in 3 Steps: (i) Persona-grounded user device context generation, (ii) Action plot generation, and (iii) Dialog utterance realization.

Plot Construction

Dialog plot is composed based on **slot-filling strategy**. Throughout the conversation, the goal for the system is trying to inquire the slot information for the target operation inten, e.g. hotel booking.

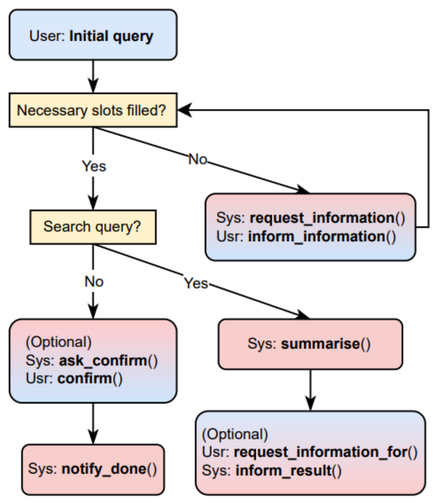


Figure 3: Plot construction for single intent dialog based on slot-filling strategy.

TOAD Dataset Stats

	MultiWoZ	PRSTO	ABCD	SGD	STAR	**TOAD**
Number of dialogs	8,438	95,671	8,034	16,142	5,820	8,087
Number of services	7	34	30	16	13	11
Total number of turns	115,434	-	177,407	329,964	127,833	37,678
Average tokens / turn	13.1	9.0	9.1	11.2	11.2	10.6
Context app interaction	×	✓	×	×	×	✓
Response style control	×	×	×	×	×	✓
Highly automatic	×	×	×	×	×	✓

Table 1: Comparison of our TOAD dataset to other existing TOD datasets.

Dialog Phenomena

Dialog Phenomena: TOAD simulates 4 complex conversation phenomena, compound, compositional, self-correction and complex referral.

Phenomena	Actions	Utterance
Compound	`restaurant_booking_reserve_table(restaurant ="French Brasserie", time="8:00 PM"), hotel_booking_search_hotel(location="Las Vegas")`	I would like to reserve a table at the French Brasserie for 8:00 PM. Also, can you help me find a hotel in Las Vegas?
Compositional	`weather_get_weather(date=get_calendar_events( name="Art Class").calendar_events_check(date).date)`	What's the weather like on the day of my Art Class event?
Self-correction	`get_movie_time(movie_name="Fast & Furious Presents: Hobbs & Shaw", location="Miami").self_correction( location="Houston")`	Could you find the showtimes for Fast & Furious Presents: Hobbs & Shaw in Miami? Actually, make that Houston instead.
Complex referral	`get_alarms(ordered_by="time", index=0).check(time)`	What's the time for my earliest alarm?

Table 2: Dialog Phenomena Examples. Actionns and utterance for initial query examples, For multi-intent phenomena, compound and compositional, we concate service name as prefix to the intent actions.

Results

NLG benchmarks

We establishes benchmarks for two response generation setups:

- Plot action to utterance
- Plot action+dialog history to utterance

Model	Act to Text						DH+Act to Text					
	Test			Zero-shot Test			Test			Zero-shot Test		
	B	R	M	B	R	M	B	R	M	B	R	M
FlanT5-250m	44.8	61.8	65.0	28.4	51.5	55.1	54.2	67.2	69.4	38.3	55.3	59.3
FlanT5-3b	**45.5**	**62.4**	**65.6**	34.2	54.3	56.9	52.8	67.5	70.9	41.9	**59.2**	**64.0**
FlanT5-11b	43.0	60.9	64.2	**35.9**	**57.8**	**60.8**	**54.9**	**68.9**	**72.0**	**44.7**	**59.2**	63.3
Llama2-7b	41.4	61.1	63.5	31.3	52.2	54.0	48.2	62.9	65.1	40.6	54.6	61.5
Llama2-13b	41.4	59.6	64.8	34.8	55.5	57.5	49.4	64.0	68.7	42.7	56.0	62.0

Table 3: Results for two benchmarks, Action to utterance and Dialog History (DH)+Action to utterance, reported in BLEU(B), Rouge-L(R) and Meteor(M) scores. All models are fine-tuned on the train set and evaluated on test and zero-shot test sets.

Experiments show that

- Dialog history can generally improve the generation performance.
- Responses with mirroring styles are easier to learn.
- More verbose responses are harder to learn.

Conclusions

In conclusion, our study explores the **naturalness** and **adaptiveness** of system responses for the next generation of TOD virtual assistants. We introduce TOAD, a dataset designed to train TOD systems for **diverse verbosity levels, mirroring styles, and realistic app context interactions**. Additionally, we present a **cost-effective and scalable automatic data generation pipeline** as a practical alternative to traditional human annotations. By addressing those critical gaps, we aim for TOAD to inspire future exploration in modeling and analyzing system response styles.