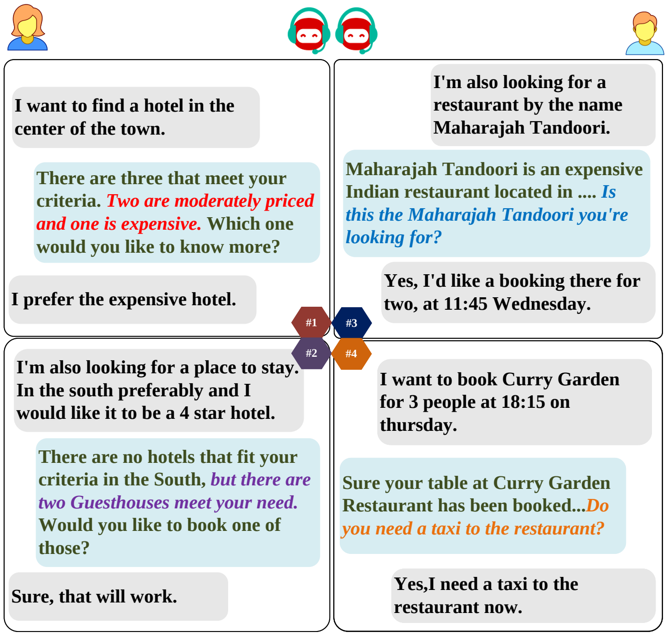
TITAN : Task-oriented Dialogues with Mixed-Initiative Interactions

Sitong Yan , **Shengli Song**[*] , **Jingyang Li** , **Shiqi Meng** and **Guangneng Hu**

School of Computer Science and Technology
Xidian University, Xi'an, China
{styan, jylee, shiqimeng}@stu.xidian.edu.cn, shlsong@xidian.edu.cn, njuhgn@gmail.com

Abstract

In multi-domain task-oriented dialogue systems, users proactively propose a series of domain-specific requests that can often be under-or over-specified, sometimes with ambiguous and cross-domain demands. System-sided initiative would be necessary to identify certain situations and appropriately interact with users to resolve them. However, most existing task-oriented dialogue systems fail to consider such mixed-initiative interaction strategies, performing low efficiency and poor collaboration ability in human-computer conversation. In this paper, we construct a multi-domain task-oriented dialogue dataset with mixed-initiative strategies **TITAN** from the large-scale dialogue corpus MultiWOZ 2.1. It contains a total of 1,800 human-human conversations where the system can either ask clarification questions actively or provides relevant information to address failure situations and implicit user requests. We report the results of several baseline models on system response generation and dialogue act prediction to assess the performance of SOTA methods on TITAN. These models can capture mixed-initiative dialogue acts, while remaining the deficiency to actively generate implicit requests and accurately provide alternative information, suggesting ample room for improvement in future studies.



I want to find a hotel in the center of the town.

There are three that meet your criteria. *Two are moderately priced and one is expensive.* Which one would you like to know more?

I prefer the expensive hotel.

I'm also looking for a place to stay. In the south preferably and I would like it to be a 4 star hotel.

There are no hotels that fit your criteria in the South, *but there are two Guesthouses meet your need.* Would you like to book one of those?

Sure, that will work.

I'm also looking for a restaurant by the name Maharajah Tandoori.

Maharajah Tandoori is an expensive Indian restaurant located in .... *Is this the Maharajah Tandoori you're looking for?*

Yes, I'd like a booking there for two, at 11:45 Wednesday.

I want to book Curry Garden for 3 people at 18:15 on thursday.

Sure your table at Curry Garden Restaurant has been booked...*Do you need a taxi to the restaurant?*

Yes,I need a taxi to the restaurant now.

Figure 1: Typical mixed-initiative interaction envolved situations. Mixed-initiative interactions can collaboratively provide optional information (part 1), proactively offer relevant information (part 2), ask clarification questions as verification for ambiguous requests(part 3), and offer a cross-domain service at an appropriate moment (part 4) that may largely benefits for commercial dialog systems.

1 Introduction

Industrial practice has focused on building task-oriented dialogue systems that can help with specific tasks such as flight reservation [Seneff and Polifroni, 2000] or bus information [Raux *et al.*, 2005] in the past decade. In a task-oriented conversation, a user converses with a system to propose a series of demands that can often be under-or over-specified, sometimes with ambiguous or cross-domain requests.

An ideal system would first identify that they were in such a situation by searching through their underlying knowledge source and then appropriately initiate interaction with either volunteer information that is not requested or ask questions of its own to become assistive and collaborative. However, existing task-oriented dialog systems fail to incorporate such **mixed-initiative** interactions (information is exchanged between user and system in turns)[Walker and Whittaker, 1990]. In this paper, we present a new research topic named mixed-initiative interaction strategies in multi-domain task-oriented dialog systems. To explore the incorporation of mixed-initiative interaction strategies in task-oriented dialogues, three research questions (RQ) need to be tackled:

- RQ 1: When the initiative should be transferred from user-side to system side?
- RQ 2: How does the system conduct appropriate initiative interactions with users?
- RQ 3: How to evaluate the effectiveness of the system-

[*] Corresponding author

sided initiative generation?

With the throughout investigation of current task-oriented dialogue dataset and considering practical conversation scenarios, four typical situations that should involve system-sided initiative is defined (RQ 1). To fill the gap in existing dialogue datasets with mixed-initiative, we present new initiative strategies to perform specific responses and construct a new dataset **TITAN** [1] with (RQ 2) for further study. The evaluation for mixed-initiative ability (RQ 3) can be explored from two separate tasks: system response generation and dialogue act prediction. Mixed-initiative interactions increase the effectiveness and efficiency of dialogue and additionally improve user experience, which suggests significance and a promising future in real, complex applications.

2 Related Work

2.1 Mixed-initiative Interactions in Open-domain Dialogue Systems

Open-domain dialogue refers to a type of conversation where user and system engage in discussions about various topics without any specific goals. Existing open-domain conversation systems generate informative responses to answer questions users request instead of asking questions or leading the conversation. In order to solve such problem, [Wu *et al.*, 2019b] creates a new dataset named DuConv where one acts as a conversation leader and the other acts as the follower and endows it with the ability of proactively and explicitly leading the conversation (introducing a new topic or maintaining the current topic).[Liu *et al.*, 2022] created a unified model that can reply to both task-oriented and open-domain requests and proposed the system-initiated transition lead dialogue mode transitions (switch from chit-chat to task-oriented or from task-oriented to chit-chat).

2.2 Mixed-initiative Interactions in Information-seeking Dialogue Systems

Conversational search is a relatively young area of research that aims at automating an information-seeking dialogue[Vakulenko *et al.*, 2021]. Most information-seeking conversations consider asking questions actively to narrow the retrieving of under-specified and ambiguous user queries[Feng *et al.*, 2020] ,[Wadhwa and Zamani, 2021],[Mass *et al.*, 2021] focus on the task of selecting the next clarification question given the conversation context by passage retrieving. [Shi *et al.*, 2022] propose a new builder system model capable of determining when to ask or execute instructions on the Minecraft Corpus Dataset. [Deng *et al.*, 2022] presents a new dataset named PACIFIC to enhance the proactivity and numerical reasoning ability of conversational question answering over hybrid contexts in finance, which introduces clarification questions from the system-side. The most relevant research is the INSCIT dataset [Wu *et al.*, 2022]. INSCIT focuses on information-seeking conversation and designs four interaction strategies: direct answer, clarification, relevant answers, and no information that can solve under or over-specified user requests. Despite great differences between open-domain conversation, information-seeking dialogue, and task-oriented dialogue systems, the design of initiative strategies and corresponding situations can provide inspiration for our work.

2.3 Mixed-initiative Interactions in Task-oriented Dialogue Systems

Task-oriented systems focus on assisting users with entity retrieving, instruction giving, and even command executing with the target of collaboration. The definition of mixed-initiative has been investigated since the early twentieth century. [Walker and Whittaker, 1995] first proposed mixed-initiative in dialogue systems and define mixed-initiative as the transfer of conversation control as the dialogue proceedings and further analyzed initiative in task-oriented and advice-giving dialogues. [Yang *et al.*, 2004] redefined initiative and control as two levels of dialogue phenomena. [Yang and Heeman, 2007] proposed that initiative normally belongs to the speaker who initiates the task in human-human conversations. Although there are differences between various studies on mixed-initiative interaction, we believe that mixed-initiative in task-oriented dialogue systems consists of volunteer information that is not asked explicitly and asks questions actively to take the lead of the conversation.

While mixed-initiative interactions have received substantial attention, how to incorporate such strategies in task-oriented dialogues remains largely unaddressed.[Balaraman *et al.*, 2020] investigated the proactivity of task-oriented in existing human-human and human-computer dialogue collections and show the deficiency of such mixed-initiative interactions in previous task-oriented dialogues. In order to simulate proactivity and construct a proactive task-oriented dialogue dataset, [Balaraman and Magnini, 2020b] simulated the capacity of task-oriented systems to provide relevant information even when not explicitly requested and exhibited proactivity strategies to offer alternative information dealing with failure situations in the restaurant retrieving task. The SimDial dataset was constructed to simulate proactive strategies and show the increase in efficiency by reducing up to 60% of dialogue turns in medium complexity. To counteract the gap of modeling dialogue as proactive behavior, [Kraus *et al.*, 2022b] has collected dialogue data through an autonomous system embedded in a serious game setting and designed four different proactive actions (None, Notification, Suggestion, Intervention) in order to serve as the user's personal advisor in a sequential planning task. The ProDial dataset was then collected online using crowdsourcing resulting in a total of 3,696 system-user exchanges.

Although recent studies have paid substantial attention to mixed-initiative interaction strategies in task-oriented dialogue systems, there remain problems largely unexplored. Our study focuses on redesigning system dialogue acts and incorporating such initiative strategies into dialogue generation. Furthermore, how to evaluate mixed-initiative strategies remains deficiencies and requires to be defined.

[1] We release TITAN dataset and code for evaluation at https://github.com/styanXDU/TITAN-evaluation-master

Metrics	SFX	WOZ2.0	FRAMES	M2M	SimDial	INSCIT	TITAN
Initiative	✗	✗	✗	✗	✓	✓	✓
Dialogues	1,006	600	1,369	**1,500**	-	250	1,440
Total turns	12,396	4,472	19,986	14,796	-	1,443	**22,372**
Total tokens	108,975	50,264	**521,876**	121,977	-	-	394,511
Avg. turns per dialogue	12.32	7.45	14.60	9.86	-	-	**15.54**
Avg. tokens per turn	8.79	11.24	12.60	8.24	-	-	**17.63**
No. of domains	-	1	2	3	1	4	**5**
No. of slot	14	4	**61**	13	25	-	30

Table 1: Comparison of our dataset to other task-oriented dialogue corpora. The numbers are provided for the training part of data except for FRAMES dataset where such division was not defined.

3 Dataset Construction

3.1 Mixed-initiative Annotation Scheme

TITAN focuses on system-sided initiatives instead of user-side. To conduct proactive dialog acts from a system appropriately, the first step is to define when is it a good moment for initiative transition from user side to system side.

According to the definition of mixed-initiative interactions in Section 1, we conclude 4 typical scenarios that embody system-sided initiatives – Specified requirement needed, Clarified requirement needed, Relevant answer needed, and Cross-domain needed. Then, to solve the problem of how to conduct a system-sided initiative at the appropriate moment, we investigate and analyze MultiWOZ2.1 meticulously and comprehensively, focusing on system acts annotation on it. Considering complex interactions and vague differences from Inform in Recommend annotation, we mainly focus on the rest four main system acts in this work. Due to the diverse utterance and coarse-grained dialog acts in the MultiWOZ dataset, we redesign 5 initiative dialog acts *RequestSelect, RequestVerify, InformAddition, NoOfferRelevant, RequestCrossDomain* and 2 no-initiative dialog acts *RequestSpecify, InformSpecific*.

3.2 Sample Dialogue Extraction

Our mixed-initiative dialogue dataset is constructed aiming at four situations presented in Section 1. Since asking questions for slot-filling and informing answers demanded are intrinsically involved in the original task-oriented dialogue system due to its fulfilling user intent target, we first investigate the situations containing failure situations in MultiWOZ 2.1 dataset, which remains largely unaddressed in existing datasets. To resolve the deficiency of dealing with over-specified situation where the system fails to search entities completely meets user goal, the dialogues with *NoOffer* system dialogue act during interaction between user and system are extracted for further revision.

Besides, we explore dialogues that proactively list optional entities with *Select* act in system turn, which occupies limited amount of primitive large-scale dataset. Consequently, we collect 1,800 representative dialogues which dominantly correspond to the situations we propose altogether. Since verification questions for ambiguous situations and cross-domain questions for new-domain requests are newly proposed situations that are not considered in MultiWOZ 2.1, we complement such active questions in the system turn annotation process on 1,800 selected dialogues.

3.3 Data Annotation Pipeline

Since the TITAN dataset focuses on system-sided initiative strategies, we separately recruit system workers for system utterances and dialog acts annotation, making two different system workers annotate system acts for each dialogue. Considering quality control in annotation, we also recruit validation workers to correct system annotations between two respective system workers. The data annotation pipeline is shown in Figure 2.

Annotation of System Response

Workers recruited are given an easy-to-operate graphical user interface to fix system responses each user-system turn. Specifically, we emphasize that the goal and user intent should remain unchanged since we consider the original dialogue as genuine user needs. Taking original dialogues as a realistic user goal, we ask system workers to rewrite the system response and remove redundant dialogue turns with certain principles during rewriting. For example, to provide an implicit request that is not asked, the system should offer information that the user asks explicitly in the following dialogue turn and remove corresponding turns. In addition, system workers should replenish clarification questions and new-domain requests at proper situations as principles.

Annotation of Mixed-initiative Stratigies

We formalize complicated dialogue act annotation as a multi-label classification task and provide specified definitions and examples for annotation workers as guidelines. In order to ensure consistency with original system acts, newly built acts can be recognized as refinement and complement of previous acts in MultiWOZ 2.1.

In particular, *Request* can be categorized into *RequestSelect* (initiative) and *RequestSpecify* (non-initiative). *Inform* can be grouped into *InformAddition* (initiative) and *InformSpecific* (non-initiative). *NoOfferRelevant* would obviously follow the original *NoOffer* act to provide relevant alternative options when the system fails to cover user request, and *RequestVerfiy* and *RequestCrossDomain* can be distinguished with their definition. Since previous system acts are rebuilt in a fine-grained categorization, we effectively avoid confusion and obscuri-

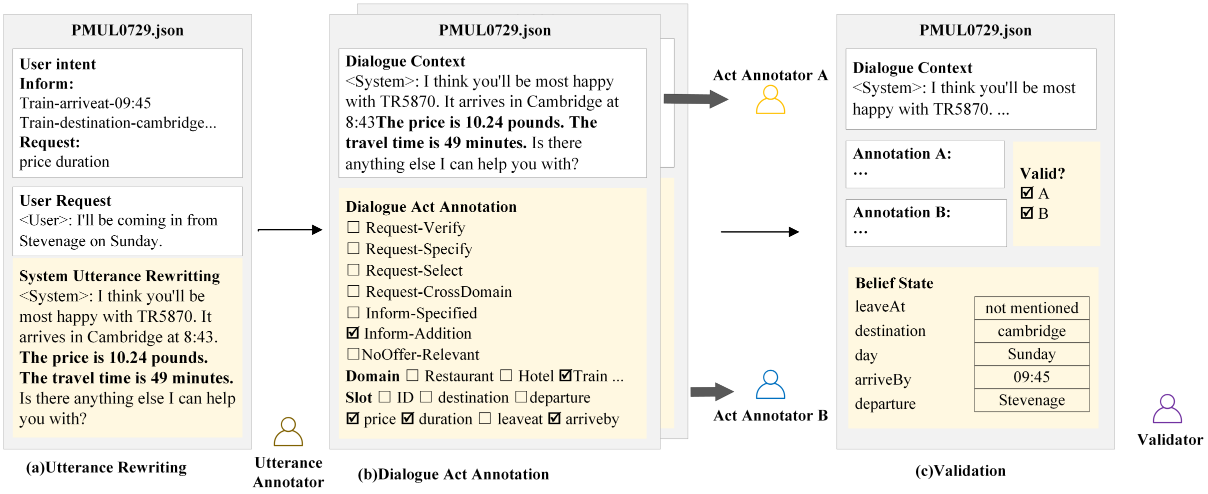

PMUL0729.json
User intent
Inform:
Train-arriveat-09:45
Train-destination-cambridge...
Request:
price duration
User Request
<User>: I'll be coming in from Stevenage on Sunday.
System Utterance Rewritting
<System>: I think you'll be most happy with TR5870. It arrives in Cambridge at 8:43. The price is 10.24 pounds. The travel time is 49 minutes. Is there anything else I can help you with?
(a)Utterance Rewriting
Utterance Annotator
PMUL0729.json
Dialogue Context
<System>: I think you'll be most happy with TR5870. It arrives in Cambridge at 8:43The price is 10.24 pounds. The travel time is 49 minutes. Is there anything else I can help you with?
Dialogue Act Annotation
☐ Request-Verify
☐ Request-Specify
☐ Request-Select
☐ Request-CrossDomain
☐ Inform-Specified
☑ Inform-Addition
☐NoOffer-Relevant
Domain ☐ Restaurant ☐ Hotel ☑Train ...
Slot ☐ ID ☐ destination ☐departure
☑ price ☑ duration ☐ leaveat ☑ arriveby
(b)Dialogue Act Annotation
Act Annotator A
Act Annotator B
PMUL0729.json
Dialogue Context
<System>: I think you'll be most happy with TR5870. ...
Annotation A:
...
Annotation B:
...
Valid?
☑ A
☑ B
Belief State
leaveAt | not mentioned
destination | cambridge
day | Sunday
arriveBy | 09:45
departure | Stevenage
Validator
(c)Validation

Figure 2: Data annotation pipeline of our dataset collection methodology. Each dialogue is annotated by system utterance rewriting, double dialogue act annotating, and validation process.

ty for annotation workers and decrease the difficulty of the annotation task.

Data Quality

Data collection is performed in a two-step process. First, all sample dialogues extracted were rewritten by system workers, and then the dialogue act annotation and validation process were launched parallelly. We recruited 48 Xidian University students including 16 undergraduates and 32 postgraduates among that 28 males and 20 females. To obtain high-quality annotations for each dialogue, we have trained the annotators and validators with detailed annotation principles before they annotate the dataset. The well-trained annotators are separated into three tasks: system response annotation, system act annotation, and validation with a ratio of 1:2:1 since the act annotation as a challenging task requires double labor to enhance the correctness.

For each dialog, there are two annotators and one validator. Given some dialog's two annotated responses, (1) Agreement: the validator reserves the response into TITAN; (2) Disagreement: (a) if one of the two annotations is considered correct by the validator, then the correct response is reserved into TITAN, (b) if both annotations are considered incorrect, then the validator acts as an annotator and her response is reserved into TITAN. In order to measure the inter-annotator agreement, system act annotation Fleiss Kappa is computed to be 0.682, suggesting a substantial agreement

4 Data Analysis

4.1 Overall Data Statistics

Following the data collection process from the previous section, a total of 28,108 user-system turns from 1,800 conversations were collected. The overall statistic compared with original dialogues is shown in Figure 3. Although we introduce verification and cross-domain request actively, the extra questions are asked at the existing dialogue turn rather than creating a new dialogue turn, which remains the length of the previous turn and avoids the inefficiency caused by the confusion. Besides, providing relevant information in failure and the implicit situation improves the efficiency of collaboration and decreases the total dialogue turn. To further explore the performance of baseline models, we split TITAN into training, dev, and test sets (shown in Table 2).

	Train	Dev	Test	Total
Convs	1,440	180	180	1,800
Turns	22,372	2,834	2,902	28,108
Tokens	394,511	49,752	50,637	494,900
Turns/Convs	15.54	15.74	16.12	15.62
Tokens/Turn	17.63	17.56	17.45	17.61

Table 2: Overall statistic of TITAN. Training, dev, and test sets are splited with 8:1:1

.

4.2 Mixed-initiative Strategies Analysis

Figure 3(c) shows the distribution of newly designed system acts annotated in our dataset. Except for general acts including bye, thank, reqmore and greeting, non-initiative strategies *InformSpecific* and *RequestSpecify* together account for around 50%, indicating the proportion for explicitly answering user requests and asking questions directly occur frequently in practice. *InformAddition* and *NoOfferRelevant* take 15.3% respectively to provide information that is not asked explicitly. *RequestCrossDomain* takes 10.6%, which is frequently aroused when users implicitly request for taxi spanning over 2 domains. *RequestSelect* is inherited from the original dialogue act and intentionally added if the user fails to specify the request after 2 human-computer turns, which occurs in a few conversations and accounts for 6.0%. Most requests can be led with explicit questions and arise from users. *RequestVerfiy* accounts for a relatively small part because verification for ambiguous situations has not been considered in the original dataset and was annotated artificially by workers, which

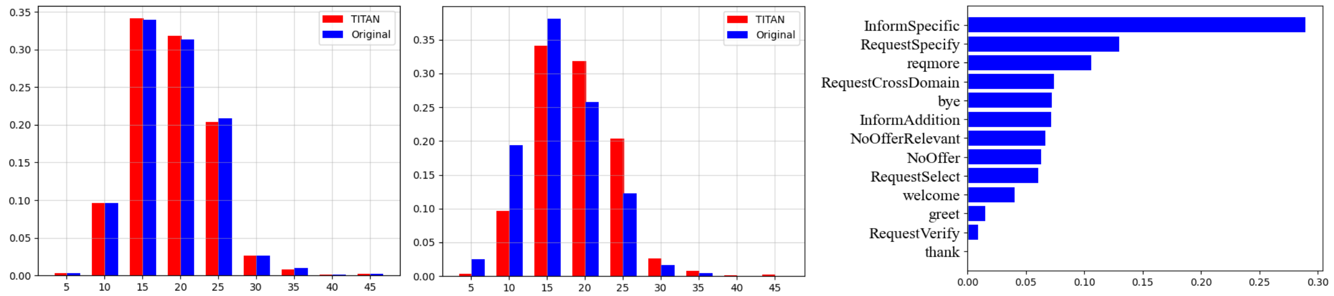


Figure 3: Dialogue length distribution (a) and distribution of the number of tokens per turn (b) in original dialogues and TITAN. Mixed-initiative dialogue acts frequency (c) in TITAN.

indicates the promising future of ambiguous clarification in task-oriented dialogue systems for further studies.

4.3 Comparision to Other Dialogue Datasets

To illustrate the contribution of our new dataset TITAN, we compare it to several important statistics with task-oriented dialogue datasets and mixed-initiative corpus most relevant to us (SFX, WOZ 2.0, FRAMES, M2M, SimDial, INSCIT). To investigate the initiative strategies in TITAN, we also report corresponding interaction acts on other conversation datasets containing relevant information or clarification questions (INSCIT[Wu *et al.*, 2022], DuConv[Wu *et al.*, 2019a], Qulac[Aliannejadi *et al.*, 2019], ShARC[Saeidi *et al.*, 2018], MultiDoc2Dial[Feng *et al.*, 2021], Abg-CoQA[Guo *et al.*, 2021], SimDial[Balaraman and Magnini, 2020a], ProDial[Kraus *et al.*, 2022a], FusedChat[Liu *et al.*, 2022].

Table 1 clearly shows that our dataset contains a more comprehensive mixed-initiative and effectively decreases interaction turns to improve the efficiency of human-computer conversation. In TITAN, we consider 5 domains inherited from MultiWOZ which possess the ability to deal with complex multi-domain, while SimDial contains single domain (*Restaurant*) with fewer slot-value pairs than our work. Furthermore, system responses in TITAN involve more sufficient semantic information and tend to generate implicit requests of entities, which leads to its larger tokens per turn average.

In Table 3, we investigate the initiative strategies in TITAN and compare them with other proactive conversation corpus both in task-oriented and open-domain dialogue. According to the definition of mixed-initiative, the strategies can be grouped into REL (provide relevant answers) and ASK (actively ask questions). Table 3 suggests that TITAN contains the most comprehensive categories of initiative strategies than other proactive dialogue corpus.

5 Evaluation

In this paper, we conduct experiments on response generation tasks for the evaluation and report several state-of-the-art baselines. Besides, we also explore the dialogue act prediction ability of these baselines in order to discuss the performance of our newly designed dialogue acts.

Dataset	REL		ASK		Task-	Information-	Open-
	Specify	Failure	Clarify	CrossDomain	Oriented	Seeking	Domain
TITAN	✓	✓	✓	✓	✓	✗	✗
INSCIT	✓	✓	✗	✗	✗	✓	✗
DuConv	✗	✗	✗	✓	✗	✗	✓
Qulac	✓	✗	✓	✗	✗	✓	✗
ShARC	✓	✗	✓	✗	✗	✓	✗
MultiDoc2Dial	✗	✗	✓	✗	✗	✓	✗
Abg-CoQA	✓	✗	✓	✗	✗	✓	✗
SimDial	✗	✓	✗	✗	✓	✗	✗
ProDial	✗	✗	✗	✗	✓	✗	✗
FusedChat	✗	✗	✗	✓	✓	✗	✓
doc2dial	✓	✗	✓	✗	✗	✓	✗
PACIFIC	✗	✗	✓	✗	✗	✓	✗

Table 3: Comparision of TITAN with existing datasets considering initiative strategies. TITAN dataset covers the most comprehensive categories of mixed-initiative strategies.

5.1 Baselines

According to the dialogue flow process, response generation models can be categorized into two groups: the context-to-text setting and the end-to-end setting [Zhao *et al.*, 2023]. Context-to-text models use the ground-truth belief states and the generated dialogue acts to carry out responses, while end-to-end models adopt the generated belief states and dialogue acts to develop responses. We choose several state-of-the-art models from both settings as baselines.

Context-to-text Baseline Models

- HDSA[Chen *et al.*, 2019]: This model proposes a multi-layer hierarchical graph to represent dialogue acts as a root-to-leaf route and uses hierarchical disentangled self-attention to model designated nodes on the dialogue act graph
- MARCO[Wang *et al.*, 2020]: This model proposes a neural co-generation the framework that generates dialogue acts and responses concurrently.
- MARCO(BERT)[Wang *et al.*, 2020]: This model is a variation of MARCO, which means MARCO is based on BERT's act prediction.
- UBAR(po)[Yang *et al.*, 2021]: This model is acquired by fine-tuning the large pre-trained unidirectional language model GPT-2 on the sequence of the entire dialog session with Policy Optimization (po) setting.

Model	MultiWOZ 2.1				TITAN			
	Inform	Success	BLEU	Combined	Inform	Success	BLEU	Combined
HDSA	86.3	70.6	**22.4**	100.8	$71.6_{\pm0.23}$	$58.1_{\pm0.39}$	**16.4**$_{\pm0.60}$	$81.2_{\pm0.55}$
MARCO	91.5	76.1	18.5	102.3	$80.3_{\pm0.31}$	$60.6_{\pm0.28}$	$14.7_{\pm0.34}$	$85.1_{\pm0.39}$
MARCO(BERT)	92.5	77.8	19.5	**104.7**	$80.7_{\pm0.19}$	$61.3_{\pm0.37}$	$14.9_{\pm0.62}$	$85.9_{\pm0.49}$
UBAR(po)	**92.7**	**81.0**	16.7	103.6	**81.3**$_{\pm0.34}$	**65.4**$_{\pm0.26}$	$14.4_{\pm0.64}$	**87.8**$_{\pm0.53}$
MultiWOZ Baseline	71.3	60.9	18.8	84.9	$60.2_{\pm0.44}$	$52.4_{\pm0.34}$	$13.9_{\pm0.43}$	$70.2_{\pm0.58}$
UniConv	72.6	62.9	**19.8**	87.6	$62.5_{\pm0.40}$	$58.7_{\pm0.38}$	**15.7**$_{\pm0.42}$	$76.3_{\pm0.41}$
LABES	76.9	63.3	17.9	88	$67.5_{\pm0.32}$	$60.3_{\pm0.35}$	$14.1_{\pm0.64}$	$78.0_{\pm0.59}$
UBAR(e2e)	**95.7**	**81.8**	16.5	**105.7**	**81.6**$_{\pm0.25}$	**64.5**$_{\pm0.38}$	$14.4_{\pm0.42}$	**87.4**$_{\pm0.36}$

Table 4: Main results of the response generation on MultiWOZ 2.1 and TITAN.

Model	F1	SER(%)				
		NoOffer	Request	Request	Request	Inform
		Relevant	Select	Verify	CrossDomain	Addition
HDSA	65.7	2.46	1.31	1.53	0.0	**1.57**
MARCO	67.1	1.75	1.07	0.62	0.0	2.14
MARCO(BERT)	68.3	1.61	0.92	**0.56**	0.0	2.08
UBAR(po)	**70.5**	**1.03**	0.76	0.53	0.0	1.97

Table 5: Results of dialogue act prediction and slot error rate under redesigned mixed-initiative with dialogue act generation methods.

End-to-end Baseline Models

- MultiWOZ Baseline[Budzianowski *et al.*, 2018]: This sequence-to-sequence model is augmented with a belief tracker and a discrete database accessing component as additional features to inform the word decisions in the decoder.
- UniConv[Le *et al.*, 2020]: This model proposes a unified neural architecture for end-to-end conversational systems, including a bi-level state tracker and a joint dialogue act and response generator.
- LABES[Zhang *et al.*, 2020]: This model proposes a probabilistic dialogue model where belief states are represented as discrete latent variables and are jointly modeled with system responses, given the user inputs
- UBAR(e2e)[Yang *et al.*, 2021]: This model is the end-to-end (e2e) setting of UBAR, which generates belief state and dialogue act for response generation.

5.2 Implementation Details

During the experiment, we first evaluate response generation tasks with several representative frameworks on TITAN and compare the results with its performance on MultiWOZ. We modify the files and codes that define ontology structures with newly designed initiative acts and retrain the models on TITAN training dataset respectively. Then, we evaluate the ability of these methods with three commonly used metrics on dev and test dataset.

To test the ability of accurately predicting initiative act and correct slot-value pairs, we then conduct dialog act prediction on context-to-text response generation baselines on test dataset, which explicitly generate dialogue act. We select two metrics that have been used in HDSA and MultiWOZ to evaluate the performance of different strategies. A single NVIDIA GeForce 3080Ti GPU with 16GB memory is used during the training and testing for both tasks.

For each model, we conduct 5 random experiments on the test set and report their mean and standard deviation. We provide the standard deviation of the evaluations in Table 4. The standard deviation of all F1 scores in Table 5 is less than 0.5.

5.3 System Response Generation

Evaluation Metrics

To evaluate the response generation ability of baseline models, we choose four widely used automatic metrics in [Sun *et al.*, 2022; Lee, 2021; He *et al.*, 2022]: **Inform** measures whether a system provides appropriate entities, **Success** assesses whether it answers all requested information, and **BLEU** is used to measure the fluency of a generated response. **Combined score**: (Inform+Success)×0.5+BLEU is as an overall quality measure as before.

Context-to-text Response Generation

Context-to-text models perform better on **Inform** metrics than end-to-end models with the usage of ground-truth belief state to generate dialogue act. In addition, Table 4 shows a pronounced decrease in **Success** metric caused by the mixed-initiative strategies we introduce failing to be sufficiently considered based on the ground-truth belief state in existing models and effects the accuracy of dialogue act generating. Here, a case study is presented to explain the low accuracy on implicit information generation in UBAR(po) in Table 6.

In the first dialogue turn, user raises the request of search for a train at certain time in Cambridge. UBAR(po) learns from training datasets and effectively generates correct dialogue act span and response span. However, in the second dialogue turn, the model fails to capture real implicit user intent and

PMUL0729.json (TITAN)

"message": ["You are planning your trip in Cambridge", "You are looking for a train. The train should arrive by **09:45** and should go to **cambridge**", "The train should leave on **sunday** and should depart from **stevenage**", "Make sure you get **train ID, price, and travel time**"] **Turn 1**

<User>: I need a train arriving in Cambridge by 09:45.

<System>: There are several that have arrival times that fit your needs. Where will you be leaving from and what day?

bspn:"<sos> [train] destination cambridge arriveat 09:45 <eos>"

aspn:" <sos> **[train][RequestSpecify] departure day** <eos>"

aspn-gen:"<sos> **[train][RequestSpecify] departure day** <eos>"

rspn-gen:"<sos> what day will you be traveling and where are you going to? <eos>"

Turn 2

<User>:I'll be coming in from Stevenage on Sunday.

...<System>: I think you'll be most happy with TR5870. It arrives in Cambridge at 8:43. **The price is 10.24 pounds. The travel time is 49 minutes.** Is there anything else I can help you with?

bspn:"<sos> [train] destination cambridge arriveat 09:45 departure Stevenage day Sunday <eos>"

aspn:"<sos> **[train][InformAddition] id arriveat price duration** <eos> "

aspn-gen:"<sos> **[train][InformAddition] id arriveat leaveat** <eos>"

rspn-gen:"<sos> sure the trainid is [value_id]. it leaves at [value_leave] and arrives at [value_arrive]. would you like to make a reservation? <eos>"

Table 6: Case Study on UBAR(po). The deficiency of inferring implicit requests of existing SOTA method largely affects the ability to volunteer implicit information.

generate dialogue act span missing the demands of the ticket price and train duration since the requests have not been asked directly, which leads to the failure of **Success** metric evaluation because **Success** metric is calculated strictly to test whether the user goals are accomplished. Such deviation indicates the limitation of existing SOTA dialogue models of inferring implicit requests of mixed-initiative strategies.

End-to-end Response Generation

In Table 4, we report four end-to-end baseline models on our dataset and compare them with their performance on MultiWOZ 2.1. All above models perform less excellently on TITAN than the original dataset, which is mainly because of the deficiency of predicting implicit user requests of existing models.

In addition, we observe that the performance of the end-to-end models on all the metrics shows a pronounced decrease compared to the context-to-text setting on our dataset. This is consistent with the general performance on MultiWOZ 2.1. Furthermore, UBAR(e2e) exhibits better performance than UBAR(po) on both datasets, which also indicates that the dialogue act generation accuracy largely affects the performance of response generation, especially the **Success** metric. The performance of all models is better shown on the MultiWOZ 2.1 dataset, which suggests the promising future of models considering mixed-initiative strategies that can enhance informative and fluent system response.

5.4 Dialogue Act Prediction

Evaluation Metrics

- Dialogue Act Prediction: For the dialogue act prediction task, we treat the domain prediction as a multi-class classification task, and the act and slot prediction as a multiple binary classifications task. The evaluation of dialogue act prediction is reported with F1 score.
- Mixed-initiative Slot Accuracy: To evaluate the accuracy of slot generation of the initiative strategies we redesign, we introduce SER (Slot Error Rate) from [Wen *et al.*, 2015] which is computed by exactly matching the slot tokens in the candidate utterances.

$$\mathrm{SER} = \frac{missing\ slot + redudant\ slot}{total\ number\ of\ slot} \tag{1}$$

Dialogue Act Analysis

As shown in Table 5, mixed-initiative strategies can be effectively learned by all baseline models. HDSA constructs a dialogue act graph based on dialogue act ontology and dialogue context to predict appropriate system acts next dialogue turn, while MARCO considers the inherent structures that are relatively important in our setting and performs better than HDSA. Benefiting from the BERT act predictor, MARCO(BERT) has improved by 1.2% on act predicting than original MARCO. UBAR incorporates the entire session of dialogue contexts to generate dialogue acts and corresponding response utterances and hence shows the best performance on both response generation and dialogue act prediction.

Mixed-initiative Strategies Analysis

In this section, we explore the mixed-initiative strategies performance on different interaction scenarios. In Table 5, RequestCrossDomain act is designed with an empty slot, so all models report correct results of **SER**. RequestVerify act has better performance than other strategies since verified slot prediction is simpler than initiative acts. NoOfferRelevant and InformAddition that are designed to resolve implicit information needed situation perform relatively poorly on **SER**. One intuitive explanation is that dialogue acts are predicted based on the ground-truth belief state, which leads to difficulties to deduce implicit information in existing dialogue models since they lack the ability of reasoning such requests and suggest ample room for further methodology.

6 Conclusion and Future Work

In conclusion, we construct TITAN, a multi-domain task-oriented dialogue dataset grounded in MultiWOZ 2.1, with redesigned mixed-initiative interaction strategies. To evaluate the quality of our dataset, we report several baselines on response generation and dialogue policy prediction that effectively learn strategies defined. In the future, we would explore the methodology that considers mixed-initiative dialogue acts to realize response generation. We would also develop automatically annotation methodology on MultiWOZ 2.1 and strive for a large-scale corpus for further studies.

Contribution Statement

Sitong Yan performed the formulation of overarching research goals, the design of mixed-initiative strategies, the development and oversight responsibility for the dataset collection and experiments planning and execution and wrote the initial draft of published work; **Shengli Song** performed the evolution of overarching research goals, the verification of the annotation methodology design, the management responsibility for the research activity planning and execution, and the revision of the published work; **Jingyang Li** performed the formulation of overarching research goals and the development of annotation methodology; **Shiqi Meng** performed the validation and analysis of dataset, conducted experiments and experiment results analysis; **Guangneng Hu** performed the verification of the experiment design and the critical review, commentary and revision of the published work.

References

[Aliannejadi *et al.*, 2019] Mohammad Aliannejadi, Hamed Zamani, Fabio A. Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019.

[Balaraman and Magnini, 2020a] Vevake Balaraman and Bernardo Magnini. Investigating proactivity in task-oriented dialogues. In *CLiC-it*, 2020.

[Balaraman and Magnini, 2020b] Vevake Balaraman and Bernardo Magnini. Proactive systems and influenceable users: Simulating proactivity in task-oriented dialogues. In *Proc. 24th Workshop Semantics Pragmatics Dialogue-Full Papers, Virually at Brandeis, Waltham, New Jersey, July (SEMDIAL)*, pages 1–12, 2020.

[Balaraman *et al.*, 2020] Vevake Balaraman, Bernardo Magnini, Fondazione Bruno Kessler, and Trento—Italy Povo. Investigating proactivity in task-oriented dialogues. *Computational Linguistics CLiC-it 2020*, page 23, 2020.

[Budzianowski *et al.*, 2018] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Chen *et al.*, 2019] Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709, Florence, Italy, July 2019. Association for Computational Linguistics.

[Deng *et al.*, 2022] Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. Pacific: Towards proactive conversational question answering over tabular and textual data in finance. *arXiv preprint arXiv:2210.08817*, 2022.

[Feng *et al.*, 2020] Song Feng, Hui Wan, Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A Lastras. doc2dial: A goal-oriented document-grounded dialogue dataset. *arXiv preprint arXiv:2011.06623*, 2020.

[Feng *et al.*, 2021] Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. Multidoc2dial: Modeling dialogues grounded in multiple documents. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

[Guo *et al.*, 2021] M. Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. Abg-coqa: Clarifying ambiguity in conversational question answering. In *Conference on Automated Knowledge Base Construction*, 2021.

[He *et al.*, 2022] Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10749–10757, 2022.

[Kraus *et al.*, 2022a] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. ProDial – an annotated proactive dialogue act corpus for conversational assistants using crowdsourcing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3164–3173, Marseille, France, June 2022. European Language Resources Association.

[Kraus *et al.*, 2022b] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. Prodial–an annotated proactive dialogue act corpus for conversational assistants using crowdsourcing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3164–3173, 2022.

[Le *et al.*, 2020] Hung Le, Doyen Sahoo, Chenghao Liu, Nancy Chen, and Steven C.H. Hoi. UniConv: A unified conversational neural architecture for multi-domain task-oriented dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1860–1877, Online, November 2020. Association for Computational Linguistics.

[Lee, 2021] Yohan Lee. 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, 2021.

[Liu *et al.*, 2022] Ye Liu, Yung-Ching Yang, Wolfgang Maier, Wolfgang Minker, and Stefan Ultes. On system-initiated transitions in a unified natural language generation model for dialogue systems, 08 2022.

[Mass *et al.*, 2021] Yosi Mass, Doron Cohen, Asaf Yehudai, and David Konopnicki. Conversational search with mixed-initiative–asking good clarification questions backed-up by passage retrieval. *arXiv preprint arXiv:2112.07308*, 2021.

[Raux *et al.*, 2005] Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. Let's go public! taking a spoken dialog system to the real world. In *in Proc. of Interspeech 2005*. Citeseer, 2005.

[Saeidi *et al.*, 2018] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. *ArXiv*, abs/1809.01494, 2018.

[Seneff and Polifroni, 2000] Stephanie Seneff and Joseph Polifroni. Dialogue management in the mercury flight reservation system. In *ANLP-NAACL 2000 Workshop: Conversational Systems*, 2000.

[Shi *et al.*, 2022] Zhengxiang Shi, Yue Feng, and Aldo Lipani. Learning to execute actions or ask clarification questions. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2060–2070, 2022.

[Sun *et al.*, 2022] Haipeng Sun, Junwei Bao, Youzheng Wu, and Xiaodong He. Bort: Back and denoising reconstruction for end-to-end task-oriented dialog. *arXiv preprint arXiv:2205.02471*, 2022.

[Vakulenko *et al.*, 2021] Svitlana Vakulenko, Evangelos Kanoulas, and Maarten De Rijke. A large-scale analysis of mixed initiative in information-seeking dialogues for conversational search. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–32, 2021.

[Wadhwa and Zamani, 2021] Somin Wadhwa and Hamed Zamani. Towards system-initiative conversational information seeking. In *DESIRES*, 2021.

[Walker and Whittaker, 1990] Marilyn Walker and Steve Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 70–78, Pittsburgh, Pennsylvania, USA, June 1990. Association for Computational Linguistics.

[Walker and Whittaker, 1995] Marilyn Walker and Steve Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. *arXiv preprint cmp-lg/9504007*, 1995.

[Wang *et al.*, 2020] Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. Multi-domain dialogue acts and response co-generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7125–7134, Online, July 2020. Association for Computational Linguistics.

[Wen *et al.*, 2015] Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[Wu *et al.*, 2019a] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy, July 2019. Association for Computational Linguistics.

[Wu *et al.*, 2019b] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. Proactive human-machine conversation with explicit conversation goals. *arXiv preprint arXiv:1906.05572*, 2019.

[Wu *et al.*, 2022] Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. Inscit: Information-seeking conversations with mixed-initiative interactions. *arXiv preprint arXiv:2207.00746*, 2022.

[Yang and Heeman, 2007] Fan Yang and Peter A Heeman. Exploring initiative strategies using computer simulation. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[Yang *et al.*, 2004] Fan Yang, Peter A Heeman, and Kristy Hollingshead. Towards understanding mixed-initiative in task-oriented dialogues. In *Eighth International Conference on Spoken Language Processing*, 2004.

[Yang *et al.*, 2021] Yunyi Yang, Yunhao Li, and Xiaojun Quan. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14230–14238, 2021.

[Zhang *et al.*, 2020] Yichi Zhang, Zhijian Ou, Min Hu, and Junlan Feng. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9207–9219, Online, November 2020. Association for Computational Linguistics.

[Zhao *et al.*, 2023] Meng Zhao, Lifang Wang, Zejun Jiang, Ronghan Li, Xinyu Lu, and Zhongtian Hu. Multi-task learning with graph attention networks for multi-domain task-oriented dialogue systems. *Knowledge-Based Systems*, 259:110069, 2023.