DialogUSR: Complex Dialogue Utterance Splitting and Reformulation for Multiple Intent Detection

Haoran Meng[1*] **Xin Zheng**[2 4*] **Tianyu Liu**[3*†] **Zizhen Wang**[3] **He Feng**[3]
Binghuai Lin[3] **Xuemin Zhao**[3] **Yunbo Cao**[3] **Zhifang Sui**[1†]
[1] MOE Key Laboratory of Computational Linguistics, Peking University, China
[2] Institute of Software, Chinese Academy of Sciences, China
[3] Tencent Cloud Xiaowei [4]University of Chinese Academy of Sciences, China
haoran@stu.pku.edu.cn;zhengxin2020@iscas.ac.cn;{rogertyliu,zizhenwang,
mobisysfeng,binghuailin,xueminzhao,yunbocao}@tencent.com;szf@pku.edu.cn

Abstract

While interacting with chatbots, users may elicit multiple intents in a single dialogue utterance. Instead of training a dedicated multi-intent detection model, we propose DialogUSR, a dialogue utterance splitting and reformulation task that first splits multi-intent user query into several single-intent sub-queries and then recovers all the coreferred and omitted information in the sub-queries. DialogUSR can serve as a plug-in and domain-agnostic module that empowers the multi-intent detection for the deployed chatbots with minimal efforts. We collect a high-quality naturally occurring dataset that covers 23 domains with a multi-step crowd-souring procedure. To benchmark the proposed dataset, we propose multiple action-based generative models that involve end-to-end and two-stage training, and conduct in-depth analyses on the pros and cons of the proposed baselines.

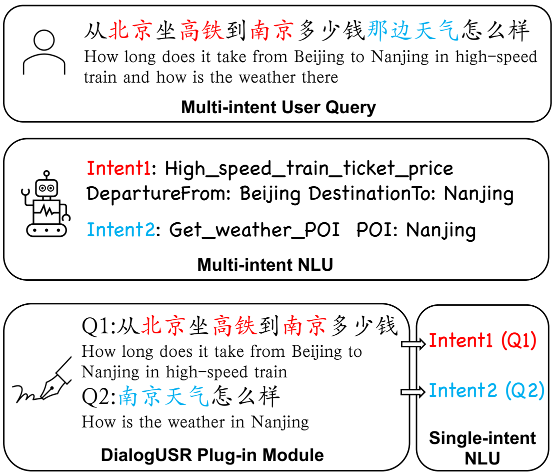

从北京坐高铁到南京多少钱那边天气怎么样
How long does it take from Beijing to Nanjing in high−speed train and how is the weather there
Multi-intent User Query
Intent1: High_speed_train_ticket_price
DepartureFrom: Beijing DestinationTo: Nanjing
Intent2: Get_weather_POI POI: Nanjing
Multi-intent NLU
Q1:从北京坐高铁到南京多少钱
How long does it take from Beijing to Nanjing in high−speed train
Q2:南京天气怎么样
How is the weather in Nanjing
DialogUSR Plug-in Module
Intent1 (Q1)
Intent2 (Q2)
Single-intent NLU

Figure 1: The task illustration for DialogUSR. It serves as a plug-in module that empowers multi-intent detection capability for deployed single-intent NLU systems.

1 Introduction

Thanks to the technological advances of natural language processing (NLP) in the last decade, modern personal virtual assistants like Apple Siri, Amazon Alexa have managed to interact with end users in a more natural and human-like way. Taking chatbots as human listeners, users may elicit multiple intents within a single query. For example, in Figure 1, a single user query triggers the inquiries on both high-speed train ticket price and the weather of destination. To handle multi-intent user queries, a straightforward solution is to train a dedicated natural language understanding (NLU) system for multi-intent detection. Rychalska et al. (2018) first adopted hierarchical structures to identify multiple user intents. Gangadharaiah and Narayanaswamy (2019) explored the joint multi-intent and slot-filling task with a recurrent neural network. Qin et al. (2020) further proposed an adaptive graph attention network to model the joint intent-slot interaction. To integrate the multi-intent detection model into a product dialogue system, the developers would make extra efforts in continuous deployment, i.e. technical support for both single-intent and multi-intent detection models, and system modifications, i.e. changes in the APIs and implementations of NLU and other related modules.

To provide an alternative way towards understanding multi-intent user queries, we propose complex dialogue utterance splitting and reformulation (DialogUSR) task with corresponding benchmark dataset that firstly splits the multi-intent query into several single-intent sub-queries and then recover the coreferred and omitted information in the sub-queries, as illustrated in Fig 1. With the proposed task and dataset, the practitioners can train a multi-intent query rewriting model that serves as a plug-in module for the existing chatbot system with minimal efforts. The trained transformation models are also domain-agnostic in the sense that the learned query splitting and rewriting skills in DialogUSR

[*] Equal contribution.
[†] Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3214–3229
December 7-11, 2022 ©2022 Association for Computational Linguistics

are generic for multi-intent complex user queries from diverse domains.

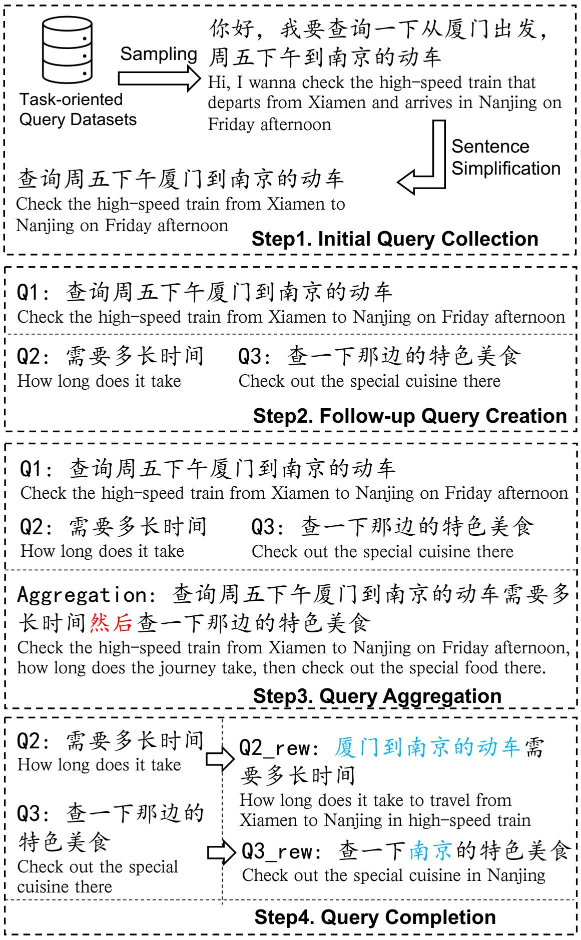
We employ a multi-step crowdsourcing procedure to annotate the dataset for DialogUSR which covers 23 domains with 11.6k instances. The naturally occurring coreferences and omissions account for 62.5% of the total human-written sub-queries, which conforms to the genuine user preferences. Specifically we first collect initial queries from 2 Chinese task-oriented NLU datasets that cover real-world user-agent interactions, then ask the annotators to write the subsequent queries as they were sending multiple intents to the chatbots, finally we aggregate the human written sub-queries and provide completed sub-queries if coreferences and omissions are involved. We also employ multiple screening and post-checking protocols in the entire data creation process, in order to ensure the high quality of the proposed dataset.

For baseline models, we carefully analyze the transformation from the input multi-intent queries to the corresponding single-intent sub-queries and summarize multiple rewriting actions, including `deletion`, `splitting`, `completion` and `causal completion` which are the local edits in the generation. Based on the summarized actions, we proposed three types of generative baselines: end-to-end, two-stage and causal two-stage models which are empowered by strong pretrained models, and conduct a series of empirical studies including the exploration on the best action combination, the model performance on different training data scale and existing multi-intent NLU datasets.

We summarize our contributions as follows[1]:

1) The biggest challenges of multi-intent detection (MID) in the deployment is the heavy code refactoring on a running dialogue system which already does a good job in single-intent detection. It motivates us to design DialogUSR, which serves as a plug-in module and eases the difficulties of incremental development.

2) Prior work on MID has higher cost of data annotation and struggles in the open-domain or domain transfer scenarios. Only NLU experts can adequately annotate the intent/slot info for a MID user query, and the outputs of MID NLU models are naturally limited by the pre-defined intent/slot ontology. In contrast, DialogUSR datasets can be easily annotated by non-experts, and the derived models are domain-agnostic in the sense that the learned query splitting, coreference/omission recovery skills are generic for distinct domains


Task-oriented Query Datasets
Sampling
Sentence Simplification
Step1. Initial Query Collection
Step2. Follow-up Query Creation
Step3. Query Aggregation
Step4. Query Completion

你好，我要查询一下从厦门出发，周五下午到南京的动车
Hi, I wanna check the high-speed train that departs from Xiamen and arrives in Nanjing on Friday afternoon

查询周五下午厦门到南京的动车
Check the high-speed train from Xiamen to Nanjing on Friday afternoon

Q1：查询周五下午厦门到南京的动车
Check the high-speed train from Xiamen to Nanjing on Friday afternoon

Q2：需要多长时间
How long does it take

Q3：查一下那边的特色美食
Check out the special cuisine there

Aggregation：查询周五下午厦门到南京的动车需要多长时间然后查一下那边的特色美食
Check the high-speed train from Xiamen to Nanjing on Friday afternoon, how long does the journey take, then check out the special food there.

Q2_rew：厦门到南京的动车需要多长时间
How long does it take to travel from Xiamen to Nanjing in high-speed train

Q3_rew：查一下南京的特色美食
Check out the special cuisine in Nanjing

Figure 2: The overview for the data collection procedure of DialogUSR. Firstly we sample initial queries from task-oriented NLU datasets (Sec. 2.1), then we hire crowdsource workers to write follow-up queries (Sec. 2.2). To aggregate the annotated queries, we propose text filler templates (marked in red, Sec. 2.3) and post-processing procedure. Finally we ask annotators to recover the missing information in the incomplete utterances (marked in blue, Sec. 2.4).

3) Presumably MID is more difficult than single intent detection (SID) given the same intent/slot ontology. From the perspective of task (re)formulation, DialogUSR is the first to convert a MID task to multiple SID tasks (the philosophy of 'divide and conquer') with a relatively low error propagation rate, providing an alternative and effective way to handle the MID task.

2 Dataset Creation

We collect a high quality dataset via a 4-step crowdsourcing procedure as illustrated in Fig 2.

[1] Code and data are provided in `https://github.com/MrZhengXin/multi_intent_2022`.

2.1 Initial Query Collection

In order to determine the topic of the multi-intent user query, we sample an initial query from two Chinese user query understanding datasets for task-oriented conversational agents, namely SMP-ECDT[2](Zhang et al., 2017) and RiSAWOZ[3] (Quan et al., 2020). Then we ask human annotators to simplify the initial queries that have excessive length (longer than 15 characters), or are too verbose or repetitive in terms of semantics[4]. RiSAWOZ is a a large-scale multi-domain Chinese Wizard-of-Oz NLU dataset with rich semantic annotations, which covers 12 domains in *tourist attraction*, *railway*, *hotel*, *restaurant*, etc. SMP-ECDT is released as the benchmark for the "domain and intent identification for user query" task in the evaluation track of Chinese Social Media Processing conference (SMP) 2017 and 2019. It covers divergent practical user queries from 30 domains which are collected from the production chatbots of iFLYTEK. We use the two source datasets as our query resources as they comprise a variety of common and naturally occurring user queries in daily life for task-oriented chatbot and cover diverse domains and topics.

2.2 Follow-up Query Creation

After specifying an initial query, we ask human annotators to put themselves in the same position of a real end user and imagine they are eliciting multiple intents in a single complex user query while interacting with conversational agents. The annotators are instructed to write up to 3 subsequent queries on what they need or what they would like to know about according to the designated initial query. Although most subsequent queries stick to the topic of the initial query, we allow the human annotators to switch to a different topic which is unrelated to the initial query[5]. For example in Figure 1, the second sub-query asks about the weather in Nanjing, where the initial query is an inquiry on the railway information. We observe that 37.3% annotated multi-intent queries involve topic switching by manually checking 300 subsampled instances in the training set, which conforms to the user behaviour in the real-world multi-intent queries.

2.3 Query Aggregation

In the pilot study, we tried to ask human annotators to manually aggregate the sub-queries but found that the derived queries are somewhat lack of variations in the conjunctions between the sub-queries, as the annotators tend to always pick up the most common Chinese conjunctions like 'and', 'or', 'then'. We even observed sloppy annotators trying to hack the annotation job by not using any conjunctions at all for each query (most queries are fluent even without conjunctions). In a nutshell, we find it challenging to screen the annotators and ensure the diversity and naturalness of the derived query in the human-only annotation. We then resort to human-in-the-loop annotation, sampling from a rich conjunction set to connect sub-queries and post-checking the sentence fluency of aggregated queries by GPT-2. After each round of annotation (we have 6 rounds of annotations), we randomly pick up 100 samples and check their quality, finding that over 95% of samples are of high quality. Actually most sentences in the Fig 9 (appendix) are fluent and natural (especially in Chinese) without cherry-picking.

More concretely we propose a set of pre-defined templates that correspond to different text infilling strategies between consecutive queries. Specifically, with a 50% chance we concatenate two consecutive queries without using any text filler. For the other 50% chance, we sample a piece of text from a set of pre-defined text fillers with different sampling weights, such as "首先" (first of all), "以及" (and), "我还想知道" (I also would like to know), "接下来" (then), "最后" (finally), and then use the sampled text filler as a conjunction while concatenating consecutive queries. Although being locally coherent, the derived multi-intent query may still exhibit some global incoherence and syntactic issues, especially for longer text. We thus post-process the derived query with a ranking procedure as an additional screening step. For each annotated query set, we generate 10 candidate multi-intent queries with different sampled templates and rank them according to language model perplexity using a GPT-2 (117M) model. We only keep the the candidate with lowest perplexity to ensure the fluency and syntactic correctness. To avoid trivial hacks in the complex query splitting, we remove

[2] `http://ir.hit.edu.cn/SMP2017-ECDT`

[3] `https://github.com/terryqj0107/RiSAWOZ`

[4] The sentence simplification phase makes the annotated multi-intent queries sound more natural, as users are not likely to elicit a lengthy query. Given the fact that we would add 2 or 3 following sub-queries to the initial queries, they should be simplified to keep a proper query length (Fig 2).

[5] In fact, we neither encourage nor discourage topic switching in the annotation instruction.

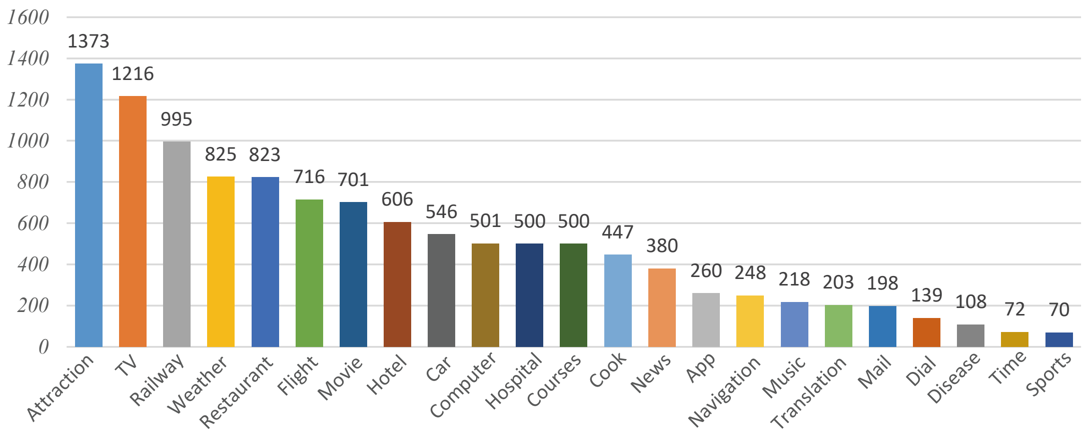


Figure 3: The domain statistics of DialogUSR, which covers diverse domains in the conversational agents.

all the punctuations in the aggregated query, which conforms to the default settings of most production chatbots, i.e. no punctuations in the spoken language understanding phase after going through the automatic speech recognition module.

2.4 Query Completion

After assembling the multi-intent user queries, we observe that incomplete utterances, such as coreferences and omissions, are frequently occurring which account for 62.5% of total human-written subsequent queries. Note that, in the annotation instruction, we do not explicitly ask the crowdsource worker to use coreferences or omissions while writing the subsequent queries in the *follow-up query creation* phase. The naturally occurring incomplete utterances reflect genuine user preferences while sending out multiple intents. To gather sufficient information while splitting multi-intent queries into independent single-intent queries, we ask another group of annotators[6] to write the completed utterances by recovering omitted and co-referred information for the incomplete queries.

2.5 Data Annotation Settings

To perform human annotation, we hired crowdsource workers from an internal data annotating group. The workers were limited to those who have abundant hand-on experiences in annotating conversational data with good records (recognized as experts in the internal assessment, rejection rate $\leq 1\%$). Additionally, all the workers were screened via a 10-case qualification test that covers various annotation tasks in Sec 2.1 to Sec 2.4 (correctly annotating 8 out of 10 cases). They were paid 0.6$ per datapoint, which is more than prevailing local minimum wage. We split the entire annotation procedure into multiple rounds and hire another group of human judges to post-check the quality of annotated dataset and filter unqualified instances after each round. In this way, we create a high-quality crowdsourcing dataset.

3 Dataset Analysis

Dataset Statistics In total, after accumulating annotations for several rounds, we obtain 11,669 instances. We conduct 6 rounds of annotation, increasing the annotation scale with each round (ranging from ∼100 instances/round to ∼4000 instances/round). On average, an aggregated multi-intent complex query from the proposed DialogUSR dataset comprises 36.7 Chinese characters by assembling 3.6 single-intent queries (including initial and follow-up queries). After recovering missing information in the query completion phase (Sec 2.4), the average lengths of completed initial query, first follow-up query, second follow-up query and third follow-up query are 11.9, 12.3, 12.4, 10.8 respectively. We split the dataset into train, validation and test sets with sizes of 10,169 , 500, 1,000 respectively.

Domain Statistics The domain statistics of DialogUSR is depicted in Fig 3. Thanks to the diverse domains of our source datasets, DialogUSR covers 23 domains that chatbot users frequently query on in their daily life. Additionally, as mentioned in

[6] The *query completion* phase starts when *follow-up query creation* phase has finished. We hire another group of annotators that did not participate in the follow-up query writing task to screen the quality of rewritten queries while doing *query completion*.

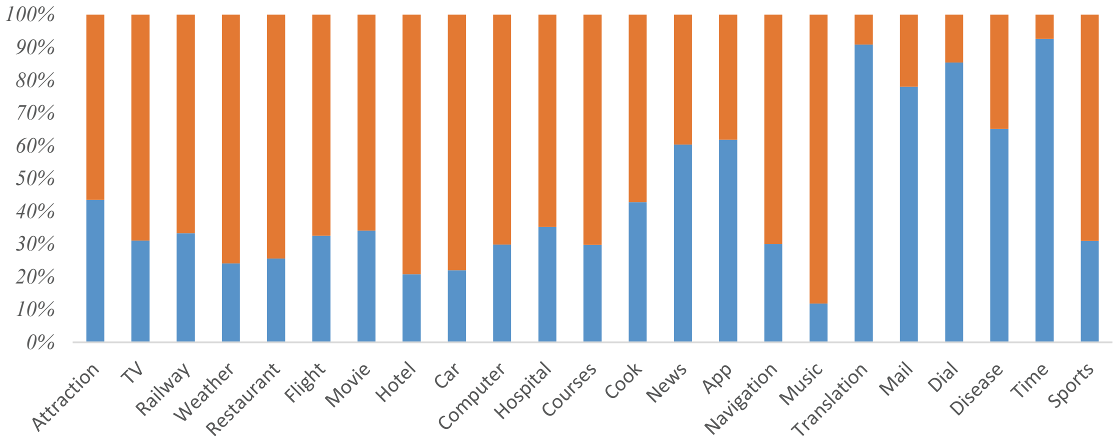


Figure 4: The ratio of the incomplete utterances in the gold outputs of DialogUSR. The blue bar signifies incomplete utterances which requires rewriting while the orange bar represents complete utterances.

Sec 2.2, the annotators proactively switch topics or domains in the data creation procedure. We find that, on average, a complex query in DialogUSR involves 1.4 domains, showing the potential usage of recognizing user intents across different domains. The models training on the DialogUSR dataset can deal with divergent situations in the practical usage while accommodating the utility of personal virtual assistant.

Incomplete Utterance Analysis Existing multi-intent detection datasets, such as MixATIS and MixSNIPS (Qin et al., 2020), were created using simple heuristic rules, e.g. adding a particular conjunction “and” while concatenating two single-intent queries. The simple heuristic datasets largely undermine the multi-intent detection in the real-world conversational agents, where users naturally interact with chatbots with coreferences and omissions. As highlighted in Sec 2.4, nearly two thirds of human-written subsequent queries are incomplete. We further show the incomplete ratio of follow-up queries for different domains in Fig 4. In the incomplete utterances, according to our statistics, only 2.4% of them belong to the coreferred phenomenon, showing that users prefer not using pronouns to refer to previously mentioned entities.

4 Baseline Models

4.1 Task Overview

As depicted in Figure 5, the input (Q1) and the output (Q4) of DialogUSR have a large text overlap. The transformation from Q1 to Q4 can be viewed as several local edits that retain the main body of the input query. We thus define several implicit actions that guide the transformation: **1)** The `Split` action (Q1→Q2) divides the complex multi-intent query into specific single-intent query with a special token. In our implementation we use the semicolon (;) and set up a heuristic rule that puts the semicolons before the text fillers if the latter appear. **2)** The `Delete` action (Q2→Q3) removes the text fillers and keep the salient queries for the subsequent actions. **3)** The `Complete` action (Q3→Q4) recovers the coreferred and omitted information in the recognized single-intent queries so that they can be effectively parsed by the existing (single-query) NLU module. **4)** The `Causal Complete` strategy consists of the `Split` action (Q1→Q2) and several `Complete` actions that echo with the token-by-token auto-regressive text generation. The difference is that `Causal Complete` strategy in DialogUSR recovers the missing information in the incomplete user utterances with a query-by-query fashion (Q5→Q6→Q7).

4.2 End-to-end Generative Models

The most straightforward way is to train a sequence-to-sequence model to learn the transformation from the multi-intent query (Q1) to the decomposed single-intent ones (Q4) in the end-to-end fashion. The models are trained to implicitly split the raw query (without punctuation) (Q1→Q2), delete the conjunctions (Q2→Q3) and recover the missing information (Q3→Q4) in one single turn of generation. Specifically given the multi-intention complex query, the model is trained to output the sequence of multiple completed independent queries “$Q_1; Q_2; ...; Q_n$; </s>”, where “;”, n, “</s>” rep-

Model	BLEU	METEOR	ROUGE	SACC	Exact Match (EM)		
					Comp.	Rew.	Avg.
End-to-end (mT5-base)	49.37	57.54	60.14	84.40	88.26	56.89	56.17
End-to-end (mT5-large)	57.07	59.83	66.47	93.50	93.53	63.04	63.90
End-to-end (mT5-xl)	64.37	62.52	**72.71**	97.90	**97.52**	70.07	71.45
End-to-end (mBART-large)	62.32	62.24	71.09	98.60	97.03	68.11	69.48
Two-stage (once,mT5-base)	49.38	58.03	60.21	83.80	89.64	56.95	56.17
Two-stage (casual,mT5-base)	55.29	60.26	65.11	85.30	89.80	61.72	60.79
Two-stage (once,mT5-large)	57.86	60.39	67.03	94.10	92.54	63.79	64.53
Two-stage (casual,mT5-large)	62.09	62.04	70.68	94.20	93.53	67.72	68.24
Two-stage (once,mT5-xl)	**64.37**	62.82	72.50	98.70	97.04	**70.33**	**71.78**
Two-stage (casual,mT5-xl)	63.73	**62.86**	72.46	**98.80**	95.07	70.28	71.62

Table 1: The benchmark for the baseline models (Fig 5). "Comp." and "Rew." correspond to the complete and rewritten (incomplete due to coreferences or omissions) queries. We report the median scores over 5 runs.

resent the query separation token, the number of queries and the end-of-sentence token, respectively.

4.3 Two-stage Generative Models

In stead of performing all three actions in one single turn, we try to guide the transformation by a step-by-step generation (Moryossef et al., 2019; Liu et al., 2019, 2021). Notably, the `Split`, `Delete` and `Complete` actions in Fig 5 can be arbitrarily permuted throughout the generation process, e.g. firstly removing text filler then split the complete the complex query (`Delete`→`Split`→`Complete`). However we observe the performance drop if we explicitly employ a 3-step generation due to the error propagation.

Two-stage model (once) we resort to a two-stage procedure that firstly splits the complex query (Q1→Q2) and then recovers the incomplete utterances (Q2→Q4). As the `Split` action is relative easy, i.e. achieving nearly 100% accuracy on the query separation, the error accumulations are largely mitigated.

Two-stage model (casual) Due to the fact that the former sub-queries would not be affected by the subsequent queries, we propose a "causal"-style query-by-query generation (Q5→Q6→Q7) in which the current sub-query to be reformulated only conditions on the prior sub-query instead of seeing the bidirectional context. Specifically, the `Causal complete` action takes place after the `Split` action. In the t-th episode of `Causal complete` action, we feed the model with incomplete queries "$q_1; ...; q_t$", and then train the model to generate the completed query Q_t. In this way, we greatly reduce the search space without the sacrifice on model performance. From an engineering standpoint, the proposed `Causal complete` action is a natural fit for the "streaming" conversational agent, i.e. simultaneous query splitting and information recovery followed by single-intent NLU while the users are eliciting multiple intents.

5 Experiment Settings

Model Setting We experiment with a variety of pretrained models via Hugging Face Transformers (Wolf et al., 2020), including mT5 (Xue et al., 2021) with three different parameter scales, namely T5-base (580M), T5-large (1.2B), T5-xl (3.7B) , and mBART-large (Liu et al., 2020b) with 340M parameters as the backbones for the end-to-end and two-stage models. They are all multi-lingual pre-trained models that support both Chinese and English DialogUSR. We use the Adam optimizer (Kingma and Ba, 2015) with the learning rate of 0.00003 and train the models for maximum 9 epochs on 4-8 A100 Gpus.

Evaluation Metrics Viewing DialohUSR as a sequence generation task, i.e. concatenating the segmented single-intent queries with semicolons like Q4 in Fig 5, we use BLEU-4 (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009), ROUGE-L (Lin, 2004), which are three commonly used automatic evaluation metrics to measure the ngram similarity with the reference in the token level. We also propose two new sentence-level metrics, namely Split Accuracy (SACC) and Exact Match (EM) to evaluate the model performance for DialogUSR. Specifically SACC measures the ratio of correct query splitting. We consider a multi-

Input(Q1): 查询周五下午厦门到南京的动车需要多长时间然后查一下那边的特色美食
Check the high-speed train from Xiamen to Nanjing on Friday afternoon, how long does the journey take, then check out the special food there.

Split(Q2): 查询周五下午厦门到南京的动车 [SP] 需要多长时间 [SP]然后查一下那边的特色美食
Check the high-speed train from Xiamen to Nanjing on Friday afternoon [SP] how long does the journey take [SP] then check out the special food there.

Delete(Q3): 查询周五下午厦门到南京的动车 [SP] 需要多长时间 [SP] 查一下那边的特色美食
Translation is the same as above

Complete(Q4): 查询周五下午厦门到南京的动车 [SP] 厦门到南京的动车需要多长时间 [SP] 查一下南京的特色美食
Check the high-speed train from Xiamen to Nanjing on Friday afternoon [SP] How long does it take to travel from Xiamen to Nanjing in high-speed train [SP] Check out the special cuisine in Nanjing

Causal Complete:
Step1(Q5): 查询周五下午厦门到南京的动车 => 查询周五下午厦门到南京的动车
Check the high-speed train from Xiamen to Nanjing on Friday afternoon => *Translation is the same as above*
Step2(Q6): 查询周五下午厦门到南京的动车 [SP] 需要多长时间 => 厦门到南京的动车需要多长时间
Check the high-speed train from Xiamen to Nanjing on Friday afternoon [SP] how long does the journey take => How long does it take to travel from Xiamen to Nanjing in high-speed train
Step3(Q7): 查询周五下午厦门到南京的动车 [SP] 需要多长时间 [SP] 然后查一下那边的特色美食 => 查一下南京的特色美食
Check the high-speed train from Xiamen to Nanjing on Friday afternoon [SP] how long does the journey take [SP] then check out the special food there => Check out the special cuisine in Nanjing

End-to-end (E2E): Q1 → Q4
Two-stage (Once): Q1 → Q2 → Q4
Two-stage (Casual): Q1 → Q2 → [Q5 → Q6 → Q7]

Figure 5: The overview for the actions taken to transform a multi-intent complex user query (Q1) to the executable single-intent queries (Q4). We use red, blue and green to highlight the text fillers, omitted information and query delimiters, respectively.

Combination	BLEU	SACC	EM
DE → (SP+CP)	46.08	77.00	51.25
(DE+CP) → SP	34.23	70.10	47.57
(SP+DE) → CP	47.60	82.00	54.15
(SP+CP) → DE	45.67	80.20	52.42
CP → (SP+DE)	45.66	78.80	52.28
SP → DE → CP	47.66	83.40	54.48
SP → (DE+CP)	**49.37**	**84.40**	**56.17**

Table 2: The exploration on the most effective action combination for the two-stage (once) model using the mT5-base models. SP, DE, CP are the abbreviations of the Split, Delete and Complete actions in Fig 5.

Model	MixSNIPS		MixATIS	
	BLEU	EM	BLEU	EM
T5-base	99.46	95.13	96.94	74.88
T5-large	99.60	97.64	98.52	88.77
T5-xl	**99.62**	**98.14**	**99.87**	**98.55**

Table 3: End-to-end model performance on the MixSNIPS and MixATIS datasets.

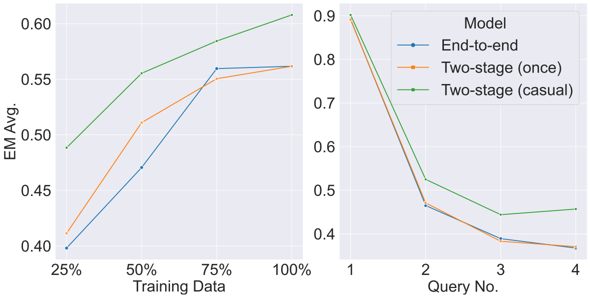


Figure 6: The model performance (mT5-base) of different training data scale (left) and sub-queries (right).

query to be correctly separated if the models split it into exactly the same number of queries as the reference:

$$\text{SACC} = \frac{1}{n} \sum_{1 \le i \le n} \mathbb{I}_{\text{len}(Q^{(i)}_{pred})=\text{len}(Q^{(i)}_{ref})},$$

where n is the number of instances, $\mathbb{I}$ is the indicator function, $Q^{(i)}_{pred}$ and $Q^{(i)}_{ref}$ are the i-th predicted and reference query list. As for EM, we consider it correct if the predicted query is exactly the same as the reference one:

$$\text{EM} = \frac{\sum_i \sum_j \mathbb{I}_{Q^{(i)}_{pred\_j}=Q^{i}_{ref\_j}}}{\sum_{1 \le i \le n} \text{len}(Q^{(i)}_{ref})},$$

where $Q^{(i)}_{pred\_j}$ and $Q^{(i)}_{ref\_j}$ represent the j-th predicted and reference query of the i-th instance. We calculate Exact Match in three different situations: EM-Complete, where we only consider the queries that does not need further modification (Q5 in Fig 5); EM-Rewritten, where delete or complete actions are needed (Q6, Q7 in Fig 5); and finally EM-Average, in which we consider all the queries.

6 Analysis and Discussions

Baseline performance Table 1 shows the performance of the baseline models on DialogUSR. For both end-to-end and two-stage generative baselines, enlarging the model parameters of mT5 models leads to a considerable performance gain, which indicates that powerful pretrained models with larger capacity are important in learning query transformation in DialogUSR. In terms of the comparisons between end-to-end and two variants of two-stage

Input: 首先查一下从广州到南宁的火车票并且我想知道有几趟列车坐火车多长时间到广州距南宁多少公里
First search the train tickets from Guangzhou to Nanning and I want to know how many trains are there; how long does it take to arrive by train; how far is it to travel from Guangzhou to Nanning

Reference:查一下从广州到南宁的火车票 [SP] 广州到南宁的火车有几趟列车 [SP] 广州到南宁坐火车多长时间到 [SP] 广州距南宁多少公里
Search the train ticket from Guangzhou to Nanning [SP] How many trains are there from Guangzhou to Nanning [SP] How long does it take to arrive in Nanning from Guangzhou by train [SP] How far is it to travel from Guangzhou to Nanning.

End-to-end:查一下从广州到南宁的火车票 [SP] 从广州到南宁有几趟列车 [SP] 从广州坐火车多长时间到 [SP] 广州距南宁多少公里
Search the train ticket from Guangzhou to Nanning [SP] How many trains are there from Guangzhou to Nanning [SP] How long does it take to travel by train from Guangzhou [SP] How far is it to travel from Guangzhou to Nanning

---

Two-stage(once):查一下从广州到南宁的火车票 [SP] 从广州到南宁有几趟列车 [SP] 从广州坐火车多长时间到南宁 [SP] 广州距南宁多少公里 [SP]
Search the train ticket from Guangzhou to Nanning [SP] How many trains are there from Guangzhou to Nanning [SP] How long does it take to arrive in Nanning by train from Nanning [SP] How far is it to travel from Guangzhou to Nanning

---

Two-stage(causal): 查一下从广州到南宁的火车票 [SP] 广州到南宁的火车有几趟列车 [SP] 广州到南宁坐火车多长时间到 [SP] 广州距南宁多少公里
Translation is the same as the reference

Figure 7: The demonstration of generated outputs for different baseline models. The query marked in red is wrong due to the missing destination of the train, while the query marked in blue is a paraphrase of the the reference.

models, we observe that for mT5-base and mT5-large, the causal-style two-stage model is the clear winner among the three models, which shows that the query-by-query transformation (Q5→Q6→Q7 in Fig 5) is the most effective way to recover the missing information while reformulating the queries. For mT5-xl, the performance gap between two-stage and end-to-end baselines is largely reduced, indicating powerful trained models may close the gap between different baselines.

We also report the model performance on the existing multi-intent detection datasets, namely MixSNIPS and MixATIS. As mentioned in Sec 3, both of them are created by inserting specific conjunctions between two complete single-intent queries from the SNIPS (Coucke et al., 2018) or ATIS (Hemphill et al., 1990) datasets, without any coreference or omission phenomenon. In other words, both of them can be effectively solved with an end-to-end model using the Delete and Split actions. The large performance gap of the same model on the MixATIS/ MixSNIPS and the proposed DialogUSR verifies that the multi-intent query splitting and reformulation task is far from solved.

Findings in the different action combinations As elaborated in Sec 4.3 and Fig 5, the Split, Delete and Complete actions can be permuted during the generation. We thus try to find the most effective action combination for the two-stage (once) model as shown in Table 2. We find that **1)** The 3-stage models[7] (SP→DE→ CP) are not necessary in the multi-stage generation compared with its two-stage variants (SP → (DE+CP)) because of the risk of error propagation (performance drop) and larger computational overhead. **2)** The Split action should be placed in the first stage, as placing it in the second stage exhibit large performance drop, e.g. SP → (DE+CP) and (DE+CP) → SP. Presumably this is because the query splitting transformation may not be robust to the potentially ill-formed rewritten queries due to the lack of exposure to the noisy training data. **3)** The Delete and Complete actions should be merged and placed in the second stage of generation. These two actions together can be viewed as a rewriting operation that deletes the conjunctions and recovers the missing information.

Detailed analysis on model outputs As the DialogUSR is actually a domain-agnostic query rewriting task, we investigate the performances of the baseline models with different training data scale in Fig 6 (left). With less training data, we observe a clear boost while employing the two-stage models. Fig 6 (right) shows the model performance while generating the sub-queries in different positions, e.g. Q5, Q6, Q7 in Fig 5) correspond to the first, second and third queries while splitting the multi-intent complex query. We observe a large performance drop while comparing the first query and the subsequent queries, because in real-world scenarios most users would not include coreferences or omissions in the query, which make it much easier to split and complete the first sub-query.

We also provide a case study for the generated outputs from different baseline models in Fig 7. Both the models trained with the two-stage strategy produce correct and executable single queries, while the end-to-end model misses the destination information in the third query, which would end up with the false parsing results in the downstream NLU modules of conversational agents.

[7] We try different actions permutations on the 3-stage models and put the most effective combination in Table 2.

7 Related Work

Incomplete Utterance Restoration To convert multi-turn incomplete dialogue into multiple single-turn complete utterance, two major paradigms are available currently. One straight-forward way is to consider it as a sequence-to-sequence problem, using models including RNN (Pan et al., 2019; Elgohary et al., 2019), Trans-PG+BERT (Hao et al., 2021) and T5 with importance token selection (Inoue et al., 2022). And since the source and target utterances are highly overlapped, another approach is to edit rather than generate from scratch, specifying the operation by sequence tagging. Pan et al. (2019) proposed Pick-and-Combine model, while Liu et al. (2020a) introduced Rewritten U-shaped Network which imitates semantic segmentation by predicting the word-level edit matrix, and with similarity Huang et al. (2021) used a semi auto-regressive generator. Later, Hao et al. (2021) proposed RUST to address the robustness issue and Jin et al. (2022) proposed hierarchical context tagging to achieve higher phrase coverage.

Multi-intent detection Spoken language understanding (SLU) which consists of intent detection and slot filling is the core in spoken dialogue systems(Tur and De Mori, 2011). Intent detection mainly aims to classify a given utterance with its intents from user inputs. Considering this strong correlation between the two tasks, some joint models are proposed based on the multi-task learning framework. (Zhang and Wang, 2016; Goo et al., 2018; Qin et al., 2019; Yao et al., 2014; Li et al., 2018). Li et al. (2018) proposed the gate mechanism to explore incorporating the intent information for slot filling. Convolutional-LSTM and capsule network have been proposed to solve the problem (Xia et al., 2018). Gangadharaiah and Narayanaswamy (2019) shows that 52% utterances are multi-intent in the Amazon internal dataset which indicate that in real world scenario, however, users often input utterance containing multi-intent. Therefore, Rychalska et al. (2018) first adopted hierarchical structures to identify multiple user intents. Qin et al. (2020) associate multi-intent detection with slots filling via graph attention network.

Larson and Leach (2022) offers a thorough overview on the existing multi-intent detection datasets. Except from MixATIS and MixSNIPS datasets, TOP (Gupta et al., 2018) contains multi-intent queries annotated in a hierarchical manner which dramatically improves the expressive power while DialogUSR contains queries and rewriting queries which can bridge the single-intent dection and multi-intent detection and also decoupling the query intent detection section and multi-intent query separation section. NLU++ (Casanueva et al., 2022) has been collected, filtered and carefully anotated by dialogue NLU experts while DialogUSR queries are created by human annotators and aggregated by rules and evaluated by model which lead to a lower cost of data annotation than NLU++.

8 Conclusion

We propose DialogUSR, a dialog utterance splitting and reformulation task and corresponding dataset, for multi-intent detection in the conversational agents. The model trained on DialogUSR can serve as a domain-agnostic and plug-in module for the existing product chatbots with minial efforts. The proposed dataset contains 11.6k high quality instances that cover 23 domains with a multi-step annotation process. We propose multiple action-based generative baselines to benchmark the dataset and analyze their pros and cons through a series of investigations.

Limitations

The proposed DialogUSR focuses on a single task for the research community and lacks of implementation details in the product conversational agents. The approaches on how the proposed DialogUSR interacts with other modules, e.g. dialog manager, ranking module for candidate NLU parsing results, remains an interesting and important research area. We position our work in the line of researches which enhances advanced conversational AI (i.e. multi-turn or multi-intent) by *query rewriting*, and leave multi-intent slot-filling entity annotation to the further work.

Acknowledgement

We thank all the anonymous reviewers for their insightful feedback. This paper is supported by the National Key Research and Development Program of China 2020AAA0106700 and the National Natural Science Foundation of China (NSFC) project U19A2065.

References

Inigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. 2022. NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1998–2013, Seattle, United States. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. 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 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 564–569, Minneapolis, Minnesota. Association for Computational Linguistics.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757, New Orleans, Louisiana. Association for Computational Linguistics.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792, Brussels, Belgium. Association for Computational Linguistics.

Jie Hao, Linfeng Song, Liwei Wang, Kun Xu, Zhaopeng Tu, and Dong Yu. 2021. RAST: Domain-robust dialogue rewriting as sequence tagging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4913–4924, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Mengzuo Huang, Feng Li, Wuhe Zou, and Weidong Zhang. 2021. Sarg: A novel semi autoregressive generator for multi-turn incomplete utterance restoration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13055–13063.

Shumpei Inoue, Tsungwei Liu, Nguyen Hong Son, and Minh-Tien Nguyen. 2022. Enhance incomplete utterance restoration by joint learning token extraction and text generation. *CoRR*, abs/2204.03958.

Lisa Jin, Linfeng Song, Lifeng Jin, Dong Yu, and Daniel Gildea. 2022. Hierarchical context tagging for utterance rewriting. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10849–10857. AAAI Press.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Stefan Larson and Kevin Leach. 2022. A survey of intent classification and slot-filling datasets for task-oriented dialog. *arXiv preprint arXiv:2207.13211*.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2–3):105–115.

Changliang Li, Liang Li, and Ji Qi. 2018. A self-attentive model with gate mechanism for spoken language understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3833, Brussels, Belgium. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020a. Incomplete utterance rewriting as semantic segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2846–2857, Online. Association for Computational Linguistics.

Tianyu Liu, Fuli Luo, Pengcheng Yang, Wei Wu, Baobao Chang, and Zhifang Sui. 2019. Towards comprehensive description generation from factual attribute-value tables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5985–5996, Florence, Italy. Association for Computational Linguistics.

Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. 2021. Towards faithfulness in open domain table-to-text generation from an entity-centric view. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13415–13423. AAAI Press.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhufeng Pan, Kun Bai, Yan Wang, Lianqiang Zhou, and Xiaojiang Liu. 2019. Improving open-domain dialogue systems via multi-turn incomplete utterance restoration. 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 1824–1833, Hong Kong, China. 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.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. 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 2078–2087, Hong Kong, China. Association for Computational Linguistics.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1807–1816, Online. Association for Computational Linguistics.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–940, Online. Association for Computational Linguistics.

Barbara Rychalska, Helena T. Glabska, and Anna Wróblewska. 2018. Multi-intent hierarchical natural language understanding for chatbots. *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 256–259.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip Yu. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3090–3099, Brussels, Belgium. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi. 2014. Spoken language understanding using long short-term memory neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 189–194.

Wei-Nan Zhang, Zhigang Chen, Wanxiang Che, Guoping Hu, and Ting Liu. 2017. The first evaluation of chinese human-computer dialogue technology. *arXiv preprint arXiv:1709.10217*.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2993–2999. IJCAI/AAAI Press.

A Implementation Detail

We run the experiments with Huggingface Transformers library on 4 Nvidia A100 GPU, with the training batch size of 96. For experiments on full training set, we set warm up step as 50. For beam search of seq2seq model, the beam size is 4. We train and test our models with 8 A100 GPUs.

Model-Size	Inference (ms)	Train (s)
End-to-end, base	71.94	289
End-to-end, large	114.95	450
End-to-end, xl	118.59	712
Two-stage(once), base	156.85	570
Two-stage(once), large	201.11	875
Two-stage(once), xl	211.26	1392
Two-stage(casual), base	173.90	1118
Two-stage(casual), large	253.06	1911
Two-stage(casual), xl	277.07	2859

Table 4: Model Inference Speed and Training Time

B Query Aggregation Detail

In Table 5, we provide conjunction probability distribution when we have four queries need to be aggregated. Conjunction0 is placed at the head of consecutive query1. Conjunction1, Conjunction2 and Conjunction3 is placed at the tail of consecutive query1, query2 and query3 respectively. As described in Table 5, Conjunction0 have 50% chance to be empty and 25% probability to be “先”(first) and another 25% chance to be “首先”(first of all). Similarly, Conjunction1 is placed at the middle of query1 and query2 with the probability described in the table and so on. Table 6 shows the probability distribution of conjunctions when three consecutive queries that need to be aggregated. We generate 10 candidate multi-intent queries by joining consecutive queries with conjunctions described in Table 5 and Table 6. After query aggregation, we calculate the perplexity of ten candidate multi-intent queries and select the most fluent sentence as multi-intent query in DialogUSR.

C DialogUSR Cases In All Domains

In Figure 8, for every twenty three domains, we respectively provide one case to show our dataset. All twenty three domains in DialogUSR are listed in Figure 8 including Attraction, TV, Railway, Weather, Restaurant, Flight, Movie, Hotel, Car, Hospital, Courses, Cook, News, App, Navigation, Music, Translation, Mail, Dial, Disease, Time, Sports. Query indicates the multi-intent query in DialogUSR and Query1 to Query4 represent the single-intent queries in DialogUSR.

	Conj0_Prob	Conj1_Prob	Conj2_Prob	Conj3_Prob
None	50%	50%	50%	50%
先	25%	0%	0%	0%
首先	25%	0%	0%	0%
然后	0%	10%	10%	10%
还有	0%	2.50%	2.50%	2.35%
我还想知道	0%	2.50%	2.50%	2.35%
另外我想知道	0%	2.50%	2.50%	2.35%
再一个就是	0%	2.50%	2.50%	2.35%
以及	0%	2.50%	2.50%	2.35%
和	0%	2.50%	2.50%	2.35%
还要	0%	2.50%	2.50%	2.35%
并且	0%	2.50%	2.50%	2.35%
再然后	0%	2.50%	2.50%	2.35%
另外	0%	2.50%	2.50%	2.35%
其次	0%	2.50%	2.50%	2.35%
同时	0%	2.50%	2.50%	2.35%
除了这个还有	0%	2.50%	2.50%	2.35%
接着	0%	2.50%	2.50%	2.35%
紧接着	0%	2.50%	2.50%	2.35%
接下来	0%	2.50%	2.50%	2.35%
最后	0%	0.00%	0.00%	2.35%

Table 5: Conjunction probability distribution in four queries cases.

	Conj0_Prob	Conj1_Prob	Conj2_Prob
None	50%	50%	50%
先	25%	0%	0%
首先	25%	0%	0%
然后	0%	10%	10%
还有	0%	2.50%	2.35%
我还想知道	0%	2.50%	2.35%
另外我想知道	0%	2.50%	2.35%
再一个就是	0%	2.50%	2.35%
以及	0%	2.50%	2.35%
和	0%	2.50%	2.35%
还要	0%	2.50%	2.35%
并且	0%	2.50%	2.35%
再然后	0%	2.50%	2.35%
另外	0%	2.50%	2.35%
其次	0%	2.50%	2.35%
同时	0%	2.50%	2.35%
除了这个还有	0%	2.50%	2.35%
接着	0%	2.50%	2.35%
紧接着	0%	2.50%	2.35%
接下来	0%	2.50%	2.35%
最后	0%	0.00%	2.35%

Table 6: Conjunction probability distribution in three queries cases.

As mentioned in Follow-up Query Creation section, we observe that 37.3% multi-intent queries involve topic switching and this phenomenon can be found in case of Translation, Time, Phone Courses etc. In Translation case, query1 to query3 is about translation while query4 What is the route to Tiananmen(去天安门的路线是什么) is about Navigation. As shown in Figure 4, a large amount of sub-queries in multi-intent query is missing information therefore they need to be rewritten by human annotators. This situation can be easily found in many cases, for example, TV case, Railway case, Weather case, Restaurant case etc. For example, in Railway case, the sub-query I would also like to know what time is the latest train? (我还想知道最晚的车次是几点？) lack the key information and human annotator rewirte the sub-query as What is the latest train number to Zhengzhou (到郑州最晚的车次是几点). We also provide a English version of twenty three cases in every domain in Fig 9.

D Broader Impact and Ethnic Consideration

Data in DialogUSR does not involve user privacy. The data source we collect from SMP-ECDT and RiSAWOZ is open source for research and is licensed under the MIT License which is a short and simple permissive license with conditions only requiring preservation of copyright and license notices.

Our generative baseline models have very low risk in terms of producing discriminatory, insulting words or divulging privacy due to the fact all the training data are strictly screened and do not include private user information or insulting content. All involved annotators voluntarily participated with decent payment.

Input	昭通到南昌的汽车最早几点发车走高速吗另外我想知道自驾要多久到 The bus from Zhaotong to Nanchang; what time does it leave at the earliest; does it take the highway; also I would like to know how long it will take to get there by self-driving
End-to-end	昭通到南昌的汽车[SP] 昭通到南昌的汽车最早几点发车[SP] 昭通到南昌的汽车走高速吗[SP] 昭通到南昌自驾要多久到 The bus from Zhaotong to Nanchang [SP] What time does the bus from Zhaotong to Nanchang leave at the earliest [SP] Does the bus from Zhaotong to Nanchang take the highway [SP] How long it will take to get there from Zhaotong to Nanchang by self-driving
Two-stage (Once)	昭通到南昌的汽车[SP] 昭通到南昌的汽车最早几点发车[SP] 昭通到南昌的汽车走高速吗[SP] 自驾昭通到南昌要多久到 The bus from Zhaotong to Nanchang [SP] What time does the bus from Zhaotong to Nanchang leave at the earliest [SP] Does the bus from Zhaotong to Nanchang take the highway [SP] How long it will take to get there by self-driving from Zhaotong to Nanchang
Two-stage (causal)	昭通到南昌的汽车[SP] 昭通到南昌的汽车最早几点发车[SP] 昭通到南昌的汽车走高速吗[SP] 自驾昭通到南昌要多久到 The bus from Zhaotong to Nanchang [SP] What time does the bus from Zhaotong to Nanchang leave at the earliest [SP] Does the bus from Zhaotong to Nanchang take the highway [SP] How long it will take to get there by self-driving from Zhaotong to Nanchang

Input	推荐一部动作电影我要看三十集左右的电视剧然后找一部90年代的动作电影 Recommend an action movie; I want to watch a TV series with about 30 episodes; and then look for an action movie in the 90s
End-to-end	推荐一部动作电影[SP] 我要看三十集左右的动作电视剧[SP] 找一部90年代的动作电影 Recommend an action movie [SP] I want to watch an action TV series with around thirty episodes [SP] Looking for an action movie from the 90s
Two-stage (Once)	推荐一部动作电影[SP] 我要看三十集左右的电视剧[SP] 找一部90年代的动作电影 Recommend an action movie [SP] I want to watch a TV series with around thirty episodes [SP] Looking for an action movie from the 90s
Two-stage (causal)	推荐一部动作电影[SP] 我要看三十集左右的动作电视剧[SP] 找一部90年代三十集左右的动作电影 Recommend an action movie [SP] I want to watch an action TV series with around thirty episodes [SP] Looking for an action movie from the 90s with around thirty episodes

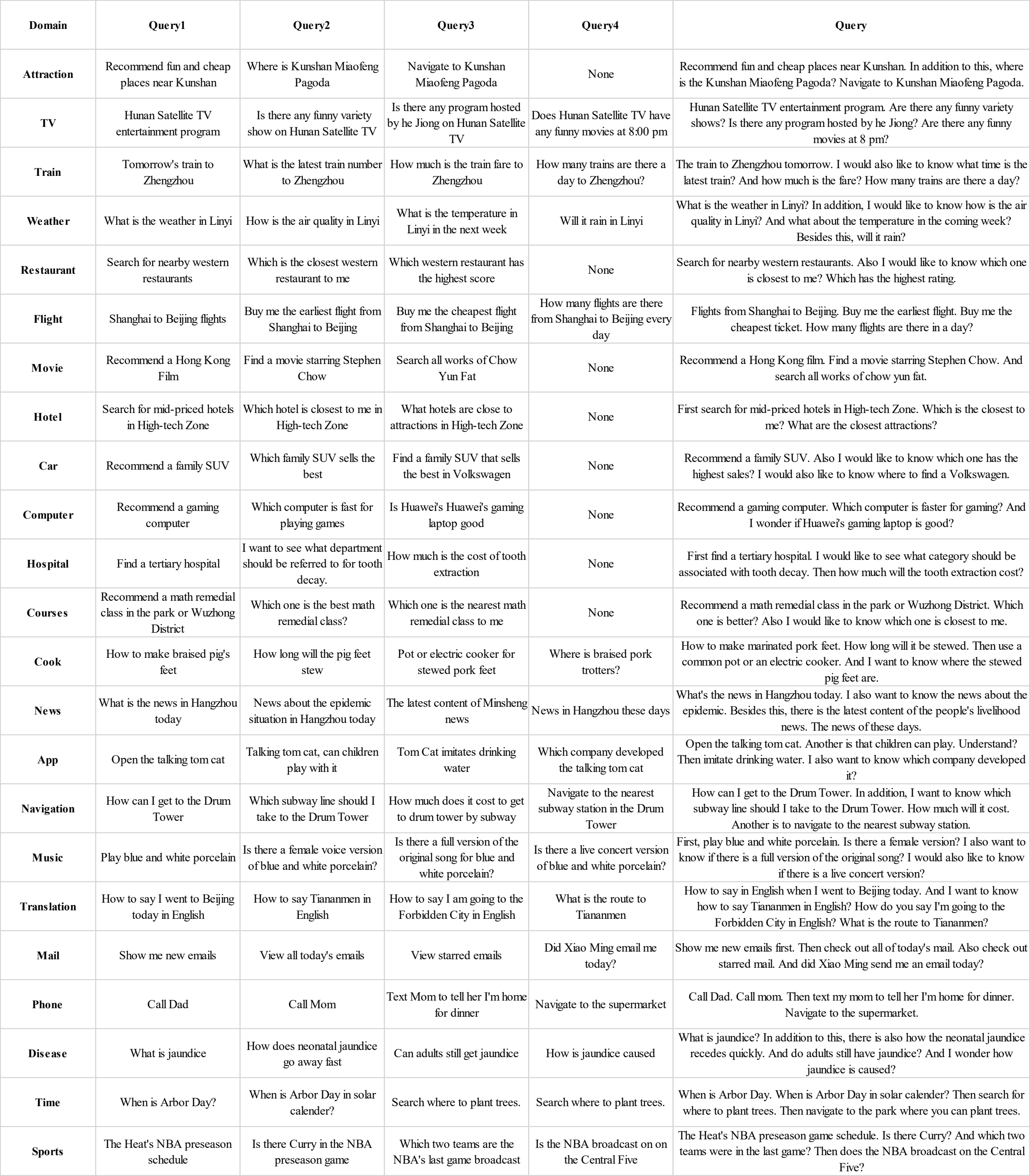
Input	首先查下下周三杭州是什么天气另外我想知道下周三有没有去杭州的航班我还想知道适合穿什么衣服 First, I want to check what weather will be in Hangzhou next Wednesday; in addition, I want to know if there is a flight to Hangzhou next Wednesday; I also want to know what clothes to wear.
End-to-end	查下下周三杭州是什么天气[SP] 下周三有没有去杭州的航班[SP] 下周三杭州的天气适合穿什么衣服 Check what weather will be in Hangzhou next Wednesday [SP] Is there a flight to Hangzhou next Wednesday [SP] What clothes should I wear for the weather in Hangzhou next Wednesday
Two-stage (Once)	查下下周三杭州是什么天气[SP] 下周三有没有去杭州的航班[SP] 下周三杭州适合穿什么衣服 Check what weather will be in Hangzhou next Wednesday [SP] Is there a flight to Hangzhou next Wednesday [SP] What clothes should I wear in Hangzhou next Wednesday
Two-stage (causal)	查下下周三杭州是什么天气[SP] 下周三有没有去杭州的航班[SP] 下周三杭州适合穿什么衣服 Check what weather will be in Hangzhou next Wednesday [SP] Is there a flight to Hangzhou next Wednesday [SP] What clothes should I wear in Hangzhou next Wednesday

Input	推荐一款越野的三厢车有没有四驱的我想要高配车 Recommend an off-road sedan; is there any four-wheel drive; I want a high-end car
End-to-end	推荐一款越野的三厢车[SP] 有没有四驱的越野的三厢车[SP] 我想要高配的越野的三厢车 Recommend an off-road sedan [SP] Is there a four-wheel drive off-road sedan [SP] I want a high-profile off-road sedan
Two-stage (Once)	推荐一款越野的三厢车[SP] 有没有四驱的越野三厢车[SP] 我想要高配的四驱越野三厢车； Recommend an off-road sedan [SP] Is there a four-wheel drive off-road sedan [SP] I want a high-end four-wheel drive off-road sedan
Two-stage (causal)	推荐一款越野的三厢车[SP] 有没有四驱的越野的三厢车[SP] 我想要高配的四驱的越野三厢车[SP] Recommend an off-road sedan [SP] Is there a four-wheel drive off-road sedan [SP] I want a high-end four-wheel drive off-road sedan

Table 7: Examples of model rewriting queries

Domain	Query1	Query2	Query3	Query4	Query
Attraction	推荐昆山附近好玩又便宜的地方	昆山妙峰塔在哪个位置	导航去昆山妙峰塔	None	推荐昆山附近好玩又便宜的地方。除了这个还有昆山妙峰塔在哪个位置？导航去昆山妙峰塔。
TV	湖南卫视娱乐节目	湖南卫视有没有搞笑的综艺节目	湖南卫视有没有何炅主持的节目	湖南卫视晚上八点有没有搞笑电影	湖南卫视娱乐节目。有没有搞笑的综艺节目？有没有何炅主持的节目？晚上八点有没有搞笑电影？
Train	明天到郑州的火车	到郑州最晚的车次是几点	到郑州的火车票价多少钱	到郑州的火车一天有几趟列车	明天到郑州的火车。我还想知道最晚的车次是几点？然后票价多少钱？还有一天有几趟列车？
Weather	临沂的天气	临沂的空气质量怎么样	临沂未来一周的温度怎么样	临沂会不会下雨	临沂的天气。另外我想知道临沂的空气质量怎么样？还有未来一周的温度怎么样？除了这个还有会不会下雨？
Restaurant	搜索附近西餐厅	离我最近的西餐厅是哪家	哪家的西餐厅评分最高	None	搜索附近西餐厅。另外我想知道离我最近的是哪家？哪家的评分最高。
Flight	上海飞往北京的航班	给我买最早的一趟上海飞往北京的航班	给我买上海飞往北京最便宜的机票	上海飞往北京一天有几趟航班	上海飞往北京的航班。给我买最早的一趟航班。给我买最便宜的机票。一天有几趟航班？
Movie	推荐一部香港电影	找一个周星驰主演的电影	搜索周润发的所有作品	None	推荐一部香港电影。找一个周星驰主演的电影。并且搜索周润发的所有作品。
Hotel	搜索高新区价格中等的酒店	高新区离我最近的酒店是哪家	高新区离景点近的酒店有哪些	None	首先搜索高新区价格中等的酒店。离我最近的是哪家？离景点近的有哪些？
Car	推荐一辆家用SUV	家用SUV销量最高的是哪一个	找一款大众销量最高的家用SUV	None	推荐一辆家用SUV。另外我想知道销量最高的是哪一个？我还想知道找一款大众的。
Computer	推荐一台游戏电脑	哪款电脑打游戏快	华为的游戏本好吗	None	推荐一台游戏电脑。哪款电脑打游戏快？并且我想知道华为的游戏本好吗？
Hospital	找一家三级医院	我想看蛀牙应该挂什么科	拔牙的费用是多少	None	首先找一家三级医院。我想看蛀牙应该挂什么科。然后拔牙的费用是多少？
Courses	推荐一个园区或者吴中区的数学辅导班	哪一家数学辅导班教的比较好	离我最近的数学辅导班是哪一个	None	推荐一个园区或者吴中区的数学辅导班。哪一家教的比较好。另外我想知道离我最近的是哪一个？
Cook	怎么做卤猪脚	猪脚要炖多久	卤猪脚用普通锅还是电饭锅	卤猪脚是哪里的美食	怎么做卤猪脚。要炖多久。然后用普通锅还是电饭锅。并且我想知道卤猪脚是哪里的美食。
News	今天杭州有什么新闻	今天杭州关于疫情的新闻	民生新闻的最新内容	杭州这几天的新闻	今天杭州有什么新闻。我还想知道关于疫情的新闻。除了这个还有民生新闻的最新内容。这几天的新闻。
App	打开会说话的汤姆猫	会说话的汤姆猫小孩子可以玩懂吗	汤姆猫模仿喝水	会说话的汤姆猫哪个公司研发的	打开会说话的汤姆猫。再一个就是小孩子可以玩懂吗？然后模仿喝水。我还想知道哪个公司研发的？
Navigation	到鼓楼怎么走	到鼓楼应该做几号线地铁	坐地铁到鼓楼要花多少钱	导航到鼓楼最近的地铁站	到鼓楼怎么走。另外我想知道应该做几号线地铁。要花多少钱。再一个就是导航到最近的地铁站。
Music	来一首青花瓷	青花瓷有没有女声版本的	青花瓷有没有完整版的原唱	青花瓷现场演唱会版本的有吗	首先来一首青花瓷。有没有女声版本的？我还想知道有没有完整版的原唱？我还想知道现场演唱会版本的有吗？
Translation	我今天去北京了英语怎么说	天安门用英语怎么说	我要去故宫用英语怎么说	去天安门的路线是什么	我今天去北京了英语怎么说。并且我想知道天安门用英语怎么说？我要去故宫用英语怎么说？去天安门的路线是什么？
Mail	给我看新邮件	查看今天的所有邮件	查看星标邮件	小明今天有给我发邮件吗	首先给我看新邮件。然后查看今天的所有邮件。另外查看星标邮件。然后小明今天有给我发邮件吗？
Dial	给爸爸打电话	给妈妈打电话	给妈妈发短信告诉她我回家吃饭	导航去超市	给爸爸打电话。给妈妈打电话。然后给妈妈发短信告诉她我回家吃饭。导航去超市。
Disease	黄疸是什么	新生儿黄疸怎么退得快	成人还会得黄疸吗	黄疸是怎么引起的	黄疸是什么？除了这个还有新生儿黄疸怎么退得快。然后成人还会的黄疸吗？并且我想知道黄疸是怎么引起的？
Time	植树节是什么时候	植树节是阳历几号	搜索哪里可以种树	导航去可以种树的公园	植树节是什么时候。植树节是阳历几号。紧接着搜索哪里可以种树。然后导航去可以种树的公园。
Sports	热火NBA季前赛的那个赛程	NBA季前赛有没有库里	NBA最近一次比赛是哪两支球队	NBA在中央五套转播吗	热火NBA季前赛的那个赛程。有没有库里？然后最近一次比赛是哪两支球队？然后NBA在中央五套转播吗？

Figure 8: DialogUSR dataset instances in all domains. Punctuations are added in the last column for better readability.



Domain	Query1	Query2	Query3	Query4	Query
Attraction	Recommend fun and cheap places near Kunshan	Where is Kunshan Miaofeng Pagoda	Navigate to Kunshan Miaofeng Pagoda	None	Recommend fun and cheap places near Kunshan. In addition to this, where is the Kunshan Miaofeng Pagoda? Navigate to Kunshan Miaofeng Pagoda.
TV	Hunan Satellite TV entertainment program	Is there any funny variety show on Hunan Satellite TV	Is there any program hosted by he Jiong on Hunan Satellite TV	Does Hunan Satellite TV have any funny movies at 8:00 pm	Hunan Satellite TV entertainment program. Are there any funny variety shows? Is there any program hosted by he Jiong? Are there any funny movies at 8 pm?
Train	Tomorrow's train to Zhengzhou	What is the latest train number to Zhengzhou	How much is the train fare to Zhengzhou	How many trains are there a day to Zhengzhou?	The train to Zhengzhou tomorrow. I would also like to know what time is the latest train? And how much is the fare? How many trains are there a day?
Weather	What is the weather in Linyi	How is the air quality in Linyi	What is the temperature in Linyi in the next week	Will it rain in Linyi	What is the weather in Linyi? In addition, I would like to know how is the air quality in Linyi? And what about the temperature in the coming week? Besides this, will it rain?
Restaurant	Search for nearby western restaurants	Which is the closest western restaurant to me	Which western restaurant has the highest score	None	Search for nearby western restaurants. Also I would like to know which one is closest to me? Which has the highest rating.
Flight	Shanghai to Beijing flights	Buy me the earliest flight from Shanghai to Beijing	Buy me the cheapest flight from Shanghai to Beijing	How many flights are there from Shanghai to Beijing every day	Flights from Shanghai to Beijing. Buy me the earliest flight. Buy me the cheapest ticket. How many flights are there in a day?
Movie	Recommend a Hong Kong Film	Find a movie starring Stephen Chow	Search all works of Chow Yun Fat	None	Recommend a Hong Kong film. Find a movie starring Stephen Chow. And search all works of chow yun fat.
Hotel	Search for mid-priced hotels in High-tech Zone	Which hotel is closest to me in High-tech Zone	What hotels are close to attractions in High-tech Zone	None	First search for mid-priced hotels in High-tech Zone. Which is the closest to me? What are the closest attractions?
Car	Recommend a family SUV	Which family SUV sells the best	Find a family SUV that sells the best in Volkswagen	None	Recommend a family SUV. Also I would like to know which one has the highest sales? I would also like to know where to find a Volkswagen.
Computer	Recommend a gaming computer	Which computer is fast for playing games	Is Huawei's Huawei's gaming laptop good	None	Recommend a gaming computer. Which computer is faster for gaming? And I wonder if Huawei's gaming laptop is good?
Hospital	Find a tertiary hospital	I want to see what department should be referred to for tooth decay.	How much is the cost of tooth extraction	None	First find a tertiary hospital. I would like to see what category should be associated with tooth decay. Then how much will the tooth extraction cost?
Courses	Recommend a math remedial class in the park or Wuzhong District	Which one is the best math remedial class?	Which one is the nearest math remedial class to me	None	Recommend a math remedial class in the park or Wuzhong District. Which one is better? Also I would like to know which one is closest to me.
Cook	How to make braised pig's feet	How long will the pig feet stew	Pot or electric cooker for stewed pork feet	Where is braised pork trotters?	How to make marinated pork feet. How long will it be stewed. Then use a common pot or an electric cooker. And I want to know where the stewed pig feet are.
News	What is the news in Hangzhou today	News about the epidemic situation in Hangzhou today	The latest content of Minsheng news	News in Hangzhou these days	What's the news in Hangzhou today. I also want to know the news about the epidemic. Besides this, there is the latest content of the people's livelihood news. The news of these days.
App	Open the talking tom cat	Talking tom cat, can children play with it	Tom Cat imitates drinking water	Which company developed the talking tom cat	Open the talking tom cat. Another is that children can play. Understand? Then imitate drinking water. I also want to know which company developed it?
Navigation	How can I get to the Drum Tower	Which subway line should I take to the Drum Tower	How much does it cost to get to drum tower by subway	Navigate to the nearest subway station in the Drum Tower	How can I get to the Drum Tower. In addition, I want to know which subway line should I take to the Drum Tower. How much will it cost. Another is to navigate to the nearest subway station.
Music	Play blue and white porcelain	Is there a female voice version of blue and white porcelain?	Is there a full version of the original song for blue and white porcelain?	Is there a live concert version of blue and white porcelain?	First, play blue and white porcelain. Is there a female version? I also want to know if there is a full version of the original song? I would also like to know if there is a live concert version?
Translation	How to say I went to Beijing today in English	How to say Tiananmen in English	How to say I am going to the Forbidden City in English	What is the route to Tiananmen	How to say in English when I went to Beijing today. And I want to know how to say Tiananmen in English? How do you say I'm going to the Forbidden City in English? What is the route to Tiananmen?
Mail	Show me new emails	View all today's emails	View starred emails	Did Xiao Ming email me today?	Show me new emails first. Then check out all of today's mail. Also check out starred mail. And did Xiao Ming send me an email today?
Phone	Call Dad	Call Mom	Text Mom to tell her I'm home for dinner	Navigate to the supermarket	Call Dad. Call mom. Then text my mom to tell her I'm home for dinner. Navigate to the supermarket.
Disease	What is jaundice	How does neonatal jaundice go away fast	Can adults still get jaundice	How is jaundice caused	What is jaundice? In addition to this, there is also how the neonatal jaundice recedes quickly. And do adults still have jaundice? And I wonder how jaundice is caused?
Time	When is Arbor Day?	When is Arbor Day in solar calender?	Search where to plant trees.	Search where to plant trees.	When is Arbor Day. When is Arbor Day in solar calender? Then search for where to plant trees. Then navigate to the park where you can plant trees.
Sports	The Heat's NBA preseason schedule	Is there Curry in the NBA preseason game	Which two teams are the NBA's last game broadcast	Is the NBA broadcast on on the Central Five	The Heat's NBA preseason game schedule. Is there Curry? And which two teams were in the last game? Then does the NBA broadcast on the Central Five?

Figure 9: English version of DialogUSR dataset instances in all domains.