

MICROSOFT DIALOGUE CHALLENGE: BUILDING END-TO-END TASK-COMPLETION DIALOGUE SYSTEMS

Xiujun Li Sarah Panda Jingjing Liu Jianfeng Gao

Microsoft, Redmond, WA, 98052, USA

ABSTRACT

This proposal introduces a Dialogue Challenge for building end-to-end task-completion dialogue systems, with the goal of encouraging the dialogue research community to collaborate and benchmark on standard datasets and unified experimental environment. In this special session, we will release human-annotated conversational data in three domains (movie-ticket booking, restaurant reservation, and taxi booking), as well as an experiment platform with built-in simulators in each domain, for training and evaluation purposes. The final submitted systems will be evaluated both in simulated setting and by human judges.

Index Terms— dialogue challenge, end-to-end task-completion dialogue

1. INTRODUCTION

There are many virtual assistants commercially available today, such as Apple’s Siri, Google’s Home, Microsoft’s Cortana, and Amazon’s Echo. With a well-designed dialogue system as an intelligent assistant, people can accomplish tasks easily via natural language interactions.

In the research community, dialogue system has been well studied for many decades. Recent advance in deep learning has also inspired the exploration of neural dialogue systems. However, it still remains a big challenge to build and evaluate multi-turn task-completion systems in a universal setting.

On one hand, conversational data for dialogue research has been scarce, due to challenges in human data collection and privacy issues. Without standard public datasets, it has been difficult for any group to build universal dialogue models that could encourage follow-up studies to benchmark upon. On the other hand, labeled datasets that are available now, while useful for evaluating partial components of a dialogue system (such as natural language understanding, dialogue state tracking), fail at end-to-end system evaluation. As a thorough evaluation of a dialogue system requires a large number of users to interact with the system at real time.

A well-adopted alternative approach is the employment of user simulators. The idiosyncratic strength and weakness of simulators for dialogue systems has been a long-standing research topic. User simulators can provide an interactive

environment for evaluating dialogue system, which is more attainable and less costly than human evaluation. The use of simulators can also foster interest and encourage research effort in exploring reinforcement learning for dialogue management.

However, the progress of dialogue research via reinforcement learning is not as fast as we have expected, largely due to the lack of a common evaluation framework, on which different research groups can jointly develop new technologies and improve their systems. In addition, the dependency on simulators often limits the scope of functionality of the implemented dialogue systems, due to the inevitable discrepancy between real users and artificial simulators. Over the past few years, we have achieved some initial success in this area. This proposal aims to further develop and mature this work and release a universal experimentation and evaluation framework by working together with research teams in the community.

In this proposal, we present a new Dialogue Challenge on “End-to-End Task-Completion Dialogue System”. This differs from previous dialogue tracks, most of which have focused on component-level evaluation. In this dialogue challenge, we will release a carefully-labeled conversational dataset in multiple domains. This data can be used by participants to develop all the modules required to build task-completion dialogue systems. We will also release an experimentation platform with built-in simulators. Each domain will have its own well-trained simulator for experimentation purpose.

In the rest of the proposal, Section 2 will provide more details about the proposed experimentation platform. Section 3 will describe the specific tasks defined in the challenge, as well as the corresponding datasets that will be released. And Section 4 will describe the final evaluation of submitted systems.

2. PLATFORM OVERVIEW

The proposed experimentation platform is illustrated in Figure 1 [1]. It consists of a user simulator (on the left) that mimics a human user and a dialogue system (on the right). In the user simulator, an agenda-based user modeling component [2] works at the dialog-act level, and controls the conversation exchange conditioned on the generated user goal, to ensure that the user behaves in a consistent, goal-oriented manner. An

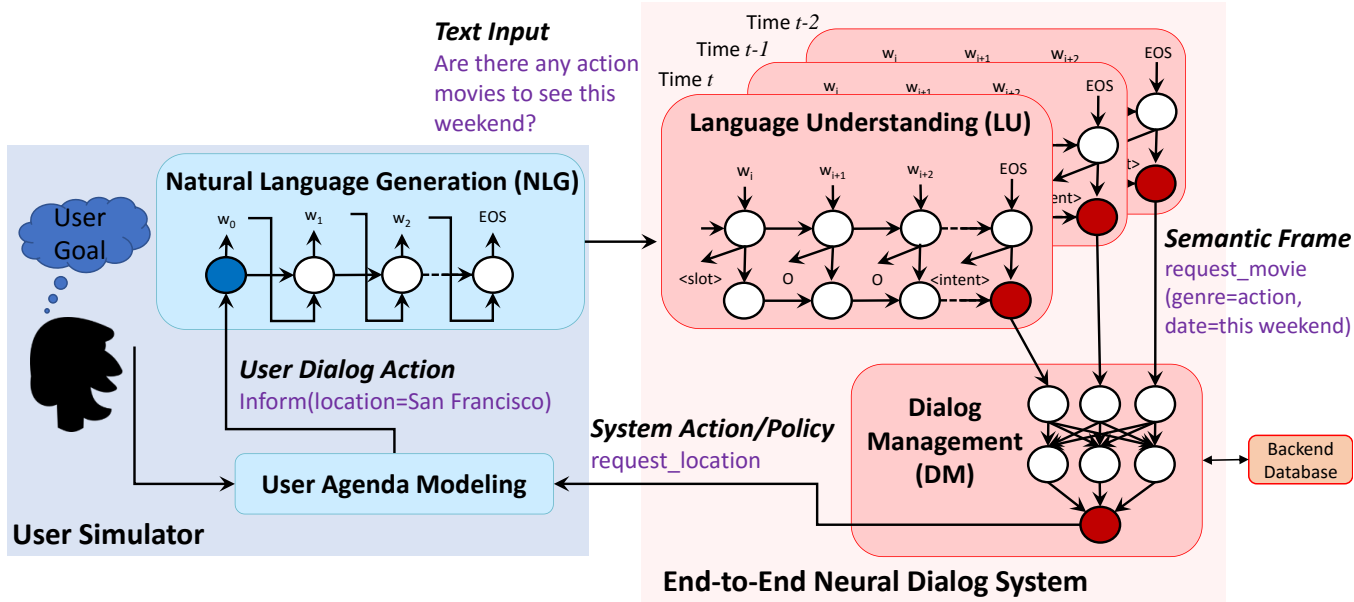


Fig. 1. Illustration of an end-to-end task-completion dialogue system.

NLU (Natural Language Understanding) module will process user’s natural language input into a semantic frame. An NLG (Natural Language Generation) module is used to generate natural language sentences responding to the user’s dialogue actions.

Although Figure 1 presents a neural dialogue system as an example, participants are free, and encouraged, to plug in any NLU/NLG modules, as long as their systems can complete a predefined task via multi-turn conversations with the user. In every turn of a conversation, the system needs to understand natural language input generated by the user or the simulator, track dialogue states during the conversation, interact with a task-specific dataset (described in Section 3), and generate an action (i.e., system response). The system action could be presented either as semantic frames (known as *dialog-acts* in simulation evaluation), or as natural language utterances generated by an NLG module.

3. TASK DESCRIPTION

In this dialogue challenge, we will release well-annotated datasets for three task-completion domains¹: movie-ticket booking, restaurant reservation, and taxi ordering. Table 1 shows the statistics of the three datasets. We will use movie ticket booking as an example to explain the specific task of building dialogue system in each domain.

¹All the datasets and code will be available at https://github.com/xiul-msr/e2e_dialog_challenge.

Task	#Intents	#Slots	#Dialogs
Movie-Ticket Booking	11	29	2890
Restaurant Reservation	11	30	4103
Taxi Ordering	11	19	3094

Table 1. The Statistics of Three Tasks.

3.1. Movie-Ticket booking task

In this task, the goal is to build a dialogue system to help users find information about movies and book movie tickets. Throughout the course of the conversation, the agent gathers information about the user’s requests, and in the end books the movie tickets or provides information about the movie in question. At the end of the conversation, the dialogue environment assesses a binary outcome (success or failure), based on (1) whether a movie ticket is booked and (2) whether the movie satisfies the user’s constraints.

3.1.1. Dataset

The data that will be released for this task was collected via Amazon Mechanical Turk. The annotation schema contains 11 intents (e.g., inform, request, confirm_question, confirm_answer, etc.), and 29 slots (e.g., moviename, starttime, theater, numberofpeople). Most of the slots are *informational* slots, which can be used to constrain the search. Others are *request* slots, with which users can request information from the agent. The final dataset to release will consist of 2890 dialogue sessions, with approximately 7.5 turns per session on average. Table 2 shows an example of annotated human-human dialogue in the movie-ticket booking task. And Table

3 shows one success and one failure dialogue example, generated by a rule-based agent and an RL agent interacting with user simulator, respectively.

3.1.2. User-Goal Set

The user goals that will be released alongside with the labeled data, are extracted from labeled dialogues by two methods. The first one extracts all the slots (known and unknown) from the first user turn (excluding the greeting turn) in each session, under the assumption that the first turn usually contains the main request from the user. The second method aggregates all the slots that appear in the user turns into one user goal. These user goals are then stored into a user-goal database for the simulator to draw from. When triggering a dialogue, the user simulator randomly samples one user goal from this database. Figure 2 shows one example user goal for the movie-ticket booking task.

```

1 New episode, user goal:
2 {
3   "request_slots": {
4     "ticket": "UNK"
5   },
6   "inform_slots": {
7     "city": "seattle",
8     "numberofpeople": "2",
9     "theater": "amc pacific place 11 theater",
10    "starttime": "9:00 pm",
11    "date": "tomorrow",
12    "moviename": "deadpool"
13  }
14 }

```

Fig. 2. An example of a user goal: the user wants to buy 2 tickets of Deadpool at 9:00 PM tomorrow at amc pacific place 11 theater, Seattle.

3.1.3. Knowledge Base

The knowledge base to be released for this task was built from labeled dialogues. The data entries are organized in JSON files, which include several KBs of size 1000 records or above. Examples of these data records are as follows:

```

1 #movie 1
2 'city': 'hamilton', 'theater': 'manville 12 plex', 'zip': '08835
3   ', 'critic_rating': 'good',
4 'date': 'tomorrow', 'state': 'NJ', 'starttime': '10:30am', '
5   genre': 'comedy', 'moviename': 'zootopia'
6
7 #movie 2
8 'city': 'seattle', 'theater': 'regal meridian 16', 'zip': '98101
9   ', 'theater_chain': 'regal meridian',
10 'state': 'WA', 'starttime': '6:30pm', 'date': 'tonight', '
11   moviename': 'zootopia'

```

Fig. 3. Data records in Movie Knowledge Base.

3.1.4. User Simulator

For the experimentation platform, we will also release a user simulator [3] for this task. The user simulator can support two formats of input:

1. *Frame-level semantics*: A dialog act form (e.g., request(moviename; genre=action; date=this weekend)) that can be used for debug purpose.
2. *Natural language*: Natural language text. To use this format, each participant needs to build their own NLU component to convert natural language input into frame-level semantics.

4. EVALUATION

To evaluate the quality of the submitted systems, we will conduct both simulation evaluation and human evaluation.

4.1. Simulation Evaluation

Three metrics will be used to measure the quality of the systems: {*success rate*, *average turns*, *average reward*}. Success rate is sometimes known as *task completion rate* – the fraction of dialogues that ended successfully. Average turns is the average length of the dialogue. Average reward is the average reward received during the conversation. There is a strong correlation among the three metrics: generally speaking, a good policy should have a high success rate, high average reward and low average turns. Here, we choose *success rate* as our major evaluation metric.

4.2. Human Evaluation

We will also conduct human evaluation for the competition. We will ask human judges to interact with the final systems submitted by participants. Besides the measurements aforementioned, each user will also give a rating on a scale of 1 to 5 based on the naturalness, coherence, and task-completion capability of the system, at the end of each dialogue session.

5. REFERENCES

- [1] Xuijun Li, Yun-Nung Chen, Lihong Li, and Jianfeng Gao, “End-to-end task-completion neural dialogue systems,” *arXiv preprint arXiv:1703.01008*, 2017.
- [2] Jost Schatzmann and Steve Young, “The hidden agenda user simulation model,” *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 733–747, 2009.
- [3] Xuijun Li, Zachary C Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen, “A user simulator for task-completion dialogues,” *arXiv preprint arXiv:1612.05688*, 2016.

Table 2. Annotated human-human dialog examples for the movie-ticket booking task and the restaurant reservation task. Left column shows natural language. Right column shows annotated dialog act form. (usr: User, agt: Agent)

Natural language	Dialog Act
Movie-Ticket Booking Task	
<p><u>usr</u>: Find me a good action movie this weekend.</p> <p><u>agt</u>: London Has Fallen is currently the number 1 action movie in America.</p> <p><u>usr</u>: Oh that sounds terrific.</p> <p><u>agt</u>: Would you like to purchase tickets to this movie?</p> <p><u>usr</u>: I would need to know what city you are in.</p> <p><u>agt</u>: Seattle</p> <p><u>usr</u>: Which theater in Seattle should I search for tickets?</p> <p><u>agt</u>: Regency or AMC please.</p> <p><u>usr</u>: Around what time do you want to go, and on which day?</p> <p><u>agt</u>: 9:30 pm any day this week.</p> <p><u>usr</u>: London Has Fallen is showing at 9:45pm on Wednesday at AMC Southcenter 16, is that showing acceptable for you?</p> <p><u>agt</u>: yes that would be great.</p> <p><u>usr</u>: Excellent, how many tickets would you like?</p> <p><u>agt</u>: I want 2 tickets please!</p>	<p><u>usr</u>: request(moviename;genre=action;date=this weekend)</p> <p><u>agt</u>: inform(moviename=london has fallen; other=number 1; genre=action)</p> <p><u>usr</u>: confirm_answer()</p> <p><u>agt</u>: request(city)</p> <p><u>usr</u>: inform(city=seattle)</p> <p><u>agt</u>: request(theater;city=seattle)</p> <p><u>usr</u>:inform(theater_chain=amc#regency)</p> <p><u>agt</u>: request(date)</p> <p><u>usr</u>: inform(starttime=9:30 pm;date=this week)</p> <p><u>agt</u>: inform(moviename=london has fallen; starttime= 9:45pm; date=wednesday; theater=amc southcenter 16)</p> <p><u>usr</u>: confirm_answer()</p> <p><u>agt</u>: request(numberofpeople)</p> <p><u>usr</u>: inform(numberofpeople=2)</p>
Restaurant Reservation	
<p><u>usr</u>: Im looking for a martini bar in Indianapolis.</p> <p><u>agt</u>: Here is the restaurant I found: High Velocity. Do you want to book?</p> <p><u>usr</u>: YES</p> <p><u>agt</u>: at what date would you like to go?</p> <p><u>usr</u>: saturday night</p> <p><u>agt</u>: at what time would you like to go?</p> <p><u>usr</u>: 8pm</p> <p><u>agt</u>: how many people are going?</p> <p><u>usr</u>: 4</p> <p><u>agt</u>: Your reservation at High Velocity for 02/27/2016 08:00PM for 4 people under Joe Does has been confirmed.</p>	<p><u>usr</u>: request(restaurantname;food=martini bar; city=Indianapolis)</p> <p><u>agt</u>: request(reservation;restaurantname=High Velocity)</p> <p><u>usr</u>: confirm_answer()</p> <p><u>agt</u>: request(date)</p> <p><u>usr</u>: inform(date=Saturday night)</p> <p><u>agt</u>: request(starttime)</p> <p><u>usr</u>: inform(starttime=8pm)</p> <p><u>agt</u>: request(numberofpeople)</p> <p><u>usr</u>: inform(numberofpeople=4)</p> <p><u>agt</u>: inform(taskcomplete;restaurantname=High Velocity; date=02/27/2016;starttime=08:00pm;numberofpeople=4; personfullname=Joe Does)</p>

Table 3. Two sample dialogs generated by a rule-based agent and an RL agent with user simulator. Left column shows scenarios where both the rule-based and RL-based agents succeed. Right column shows scenarios where the rule-based agent fails, while the RL agent succeeds.(usr: User, agt: Agent)

<pre>New episode, user goal: { "request_slots": { "ticket": "UNK" }, "diaact": "request", "inform_slots": { "city": "seattle", "numberofpeople": "2", "theater": "regal meridian 16", "starttime": "9:25 pm", "date": "tomorrow", "moviename": "zoolander 2" } }</pre>	<pre>New episode, user goal: { "request_slots": { "ticket": "UNK" "theater": "UNK" "starttime": "UNK" }, "diaact": "request", "inform_slots": { "numberofpeople": "3", "date": "tomorrow", "moviename": "10 cloverfield lane" } }</pre>
Rule-based agent	Rule-based Agent
<p>usr: Can I get some tickets for zoolander 2? agt: What movie are you interested in? usr: I want to watch zoolander 2. agt: What time would you like to see it? usr: I want to watch at 9:25 pm. agt: What city you would like? usr: I want to watch at seattle. agt: What date would you like to watch it? usr: I want to set it up tomorrow. agt: Which theater would you like? usr: I want to watch at regal meridian 16. agt: How many tickets do you need? usr: I want 2 tickets please! agt: Great - I was able to purchase 2 tickets for you to see zoolander 2 tomorrow at regal meridian 16 theater in seattle at 9:25 pm. usr: Thank you. agt: Thank you.</p>	<p>usr: What is the start time for 10 cloverfield lane? agt: What movie are you interested in? usr: I want to watch at 10 cloverfield lane. agt: What time would you like to see it? usr: What time is it playing tomorrow? I need 3 tickets. agt: What city you would like? usr: I do not care. agt: What date would you like to watch it? usr: I want to set it up tomorrow. agt: Which theater would you like? usr: Which theater is available? agt: How many tickets do you need? usr: I want 3 tickets please! agt: Okay, I have booked 3 tickets for 10 cloverfield lane tomorrow. usr: Thank you. agt: Thank you.</p>
Success	Failure: Agent does not answer all the questions (starttime, theater) from user.
RL agent	RL Agent
<p>usr: Can I get tickets for zoolander 2 tomorrow? agt: Which city would you like? usr: I want to watch at seattle. agt: How many tickets do you need? usr: I want 2 tickets please! agt: 9:25 pm is available. usr: I want to watch at regal meridian 16. agt: Great - I was able to purchase 2 tickets for you to see zoolander 2 tomorrow at regal meridian 16 theater in seattle at 9:25 pm. usr: Thank you. agt: Thank you.</p>	<p>usr: Which theater can I book 3 tickets for 10 cloverfield lane? agt: What time would you like to see it? usr: Which theater and start time are available tomorrow? agt: 11:45am is available. usr: Which theater is available? agt: regal la live stadium 14 is available. usr: Could you help me to book the tickets? agt: Okay - I was able to book 3 tickets for you to see 10 cloverfield lane at regal la live stadium 14 at 11:45am tomorrow. usr: Thank you agt: Thank you</p>
Success	Success