MULTI-DOMAIN TASK-ORIENTED DIALOG CHALLENGE II

DSTC-9 TRACK PROPOSAL

1 Introduction

There has been an increasing interest in building a dialog system crossing multiple domains to accomplish a complex goal [1, 2, 3, 4, 5, 6]. With the success of Multi-Domain Task Completion Dialog Challenge in DSTC-8 Track 1, we continue with the effort of building dialog systems under the multi-domain setting in this proposal. Compared with the previous challenge, in this track, we extend the tasks by incorporating new datasets, creating new sub-tasks, and providing a new development platform. We specifically focus on two aspects of dialog systems: language portability and end-to-end system complexity.

First, with the rapid globalization process, the need for building dialog systems that supports multiple languages in the same or a similar scenario is ever increasing. However, building a task-oriented dialog system requires a considerable amount of annotated training data, and for some languages, the training corpus is very limited. Therefore, cross-lingual transfer learning becomes a popular topic during the years and helps reduce the cost of developing dialog systems for resource-poor target languages. To advance the state-of-the-art cross-lingual technology in building dialog systems, we introduce the task of cross-lingual dialog state tracking in this track.

Second, with the advancement of end-to-end learning, building a dialog system, and evaluating its performance in an end-to-end fashion has attracted increasing interest from the community. In this challenge, by offering the end-to-end dialog system task, we also provide the latest development platform that enables users to build, evaluate, and diagnose a full dialog system with ease.

2 Task Description

This track consists of two tasks:

- Participants will build a cross-lingual multi-domain dialog state tracker.
- Participants will develop multi-domain end-to-end dialog systems with the provided development platform.

2.1 Multi-domain Dialog State Tracking

Dialog state tracking (DST) is the process of updating the dialog state at each turn summarizing the entire conversation until the current turn. Since most dialog systems predict the next dialog acts based on the tracked information of dialog states, DST is viewed as one of the key components in building a dialog system. In the history of DSTC, a series of tasks have been introduced to foster the development of dialog state tracker. From DSTC-1 to DSTC-4, mono-lingual DST tasks were introduced based on datasets ranging from machine-to-machine, human-to-machine, to human-to-human settings [7, 8, 9, 10]. In DSTC-5 [11], a cross-language dialog state tracking task was introduced to address the problem of adaptation to a new language, seeking to build a tracker for the target language using resources in the source language and the corresponding machine-translated sentences in the target language. DSTC-8 [5] introduced multi-domain DST utilizing domain and slot description to learn the semantic meaning of slots to support scaling to unseen APIs.

In this task, following a similar scheme as in DSTC-5, our goal is to build a cross-lingual dialog state tracker with a training set in the rich resource language and a small development set in the low resource language. The performance of each dialog state tracker will be evaluated on an unlabelled test set in the low resource language and compared with reference annotation. In particular, we will offer two sub-tasks based on MultiWOZ 2.1 [12] and CrossWOZ [13], respectively.

- MultiWOZ 2.1. MultiWOZ is a multi-domain dialog dataset spanning 7 distinct domains and containing over 10,000 dialogs under the travel planning setting. We augment the dataset with Chinese translations as the development/test set. English is considered as the rich resource language and Chinese as the low resource dataset.
- CrossWOZ. CrossWOZ is the first large scale Chinese multi-domain task-oriented dialog dataset spanning 5 distinct domains and containing over 6,000 dialogs under the setting of travel planning to Beijing. English translation is provided along with the original Chinese corpus. In this sub-task, Chinese is considered as the rich resource language and English as the low resource dataset.

We evaluate the performance of the dialog state tracker using two metrics:

- Joint Goal Accuracy. This metric evaluates whether the predicted dialog state is exactly equal to the ground truth.
- Slot Precision/Recall/F1. These metrics evaluate whether the predicted labels for individual slots in dialog state are equal to the ground truth, microaveraged over all slots.

The final ranking will be solely based on joint goal accuracy.

2.2 Multi-domain End-to-end Dialog Challenge

Most prior works focus on improving individual components in a dialog system, without evaluating the performance of the entire system. However, the modular performance improvement does not necessarily contribute to the end-to-end performance increment. To foster the development of end-to-end dialog systems, DSTC-8 Track 1 introduced an end-to-end multi-domain task (Task 1) and received reasonable submissions from participants.

In this task, we continue with the same setting as DSTC-8 Track 1 Task 1 with two changes:

- Dataset. Instead of using MultiWOZ 2.0 [4] as in DSTC-8, we will build dialog systems based on MultiWOZ 2.1 [12] in this task. Compared with the previous version, MultiWOZ 2.1 re-annotated states and utterances based on the original utterance to fix the original noisy annotation. It also contains user dialog act annotation, which is missing in MultiWOZ 2.0. We believe that with MultiWOZ 2.1, participants are empowered to build more effective dialog systems.
- Development Platform. ConvLab [14] is employed as the development platform in the previous challenge. In this task, we are providing the next generation of ConvLab (ConvLab-2), which integrates more powerful models and supports more datasets on top of ConvLab. Also, it includes an interactive visual toolkit that enables searchers to inspect the components and diagnose the dialog systems with an enhanced evaluator.

The participants are encouraged to experiment with various approaches based on ConvLab-2 to build a dialog system that takes natural language utterance as input, tracks dialog states during the conversation, interacts with a task-specific dataset, and generates a system response at each turn. Both automatic evaluation and human evaluation results will be reported in the challenge.

- Automatic Evaluation. We will provide an automatic evaluation script equipped with an end-to-end user simulator and evaluator, and report a range of metrics including task success rate, return (reward), number of turns for dialog policy, book rate, and precision/recall/F1 score for intent/slot detection.
- Human Evaluation. For the human evaluation, we will crowdsource the work on Amazon Mechanic Turk so human judges can communicate with the agent via natural language, and make a judgment of the system based on the dialog success/failure, language understanding score, and response appropriateness score.

The final ranking only considers the success rate in human evaluation results.

3 Resources

We will provide the following resources during the challenge.

3.1 ConvLab-2

We will provide ConvLab-2, the next generation of dialog development platform built based on ConvLab, to facilitate participants' development efficiency. ConvLab-2 inherits the framework and models from ConvLab and incorporates new features, including most recent state-of-the-art models, an analysis tool, an interactive tool, etc. ConvLab-2 serves as the following functionalities:

- Toolkit for building dialog systems with both conventional pipeline approaches and end-to-end approaches. The platform includes state-of-the-art models for NLU, dialog state tracker, policy, NLG, and end-to-end models. The interfaces between modules, knowledgebase, and backend systems are designed to support multiple datasets so that APIs calls to the knowledge base can be easily made.
- Toolkit for dialog system evaluation using both automatic evaluation and human evaluations. For automatic evaluation, it consists of end-to-end user simulators and evaluators for component-wise modules and end-to-end dialog systems. For human evaluation, the tools to interact with human judges on Amazon Mechanic Turk are provided.
- Toolkit for system diagnosis. It consists of an interactive interface that not only illustrates the output of each module but also enables users to modify the result and diagnose the end-to-end performance with new outputs. It also consists of an analysis tool that considers statistics extracted from the conversations between the user simulator and the dialog system.

3.2 Datasets

Both datasets of MultiWOZ 2.1 and CrossWOZ are incorporated into ConvLab-2 in its original language. The interface to interact with these datasets are fully supported in ConvLab-2, and multiple trained models associated with the datasets will be provided.

In the development stage of the challenge, we will translate partial datasets of MultiWOZ 2.1 and CrossWOZ to Chinese and English, respectively, and release them to the participants as the development set. In the test stage, unlabelled test sets will be released to the participants in the same language as in the development stage.

4 Dialog Examples

Please see an example dialog for CrossWOZ in Table 1 and MultiWOZ 2.1 in Table 2, respectively.

5 Organizers

Baolin Peng, Jianfeng Gao, Jinchao Li, Lars Liden, Minlie Huang, Qi Zhu, Runze Liang, Ryuichi Takanobu, Shahin Shayandeh, Swadheen Shukla, Zheng Zhang
Contact information:
Jinchao Li (jincli@microsoft.com), Qi Zhu (zhu-q18@mails.tsinghua.edu.cn)

References

[1] Jianfeng Gao, Michel Galley, Lihong Li, et al. Neural approaches to conversational ai. volume 13, pages 127–298. Now Publishers, Inc., 2019.

[2] Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[3] Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proc. Meeting on Discourse and Dialogue (SIGDIAL)*, 2017.

[4] Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5016–5026, 2018.

[5] Seokhwan Kim, Michel Galley, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, et al. The eighth dialog system technology challenge. *arXiv preprint arXiv:1911.06394*, 2019.

[6] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *arXiv preprint arXiv:1909.05855*, 2019.

[7] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, 2013.

[8] Matthew Henderson, Blaise Thomson, and Jason Williams. The second dialog state tracking challenge. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 263, 2014.

[9] Matthew Henderson, Blaise Thomson, and Jason D Williams. The third dialog state tracking challenge. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 324–329. IEEE, 2014.

[10] Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, and Matthew Henderson. The fourth dialog state tracking challenge. In *Dialogues with Social Robots*, pages 435–449. Springer, 2017.

[11] Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, Matthew Henderson, and Koichiro Yoshino. The fifth dialog state tracking challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 511–517. IEEE, 2016.

[12] Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tür. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *CoRR*, abs/1907.01669, 2019.

[13] Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 2020.

[14] Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Zheng Zhang, Yaoqin Zhang, Xiang Li, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, and Jianfeng Gao. ConvLab: Multi-domain end-to-end dialog system platform. In *Proc. Conf. Association for Computational Linguistics (ACL): System Demonstrations*, July 2019.

Table 1: An example of partial dialog for CrossWOZ dataset.

Speaker	Utterance	Dialog States
User	你好，可以帮我找一个人均消费是100-150元，评分是5分的餐馆吗? Hello, can you help me find a restaurant with a per capita consumption of 100-150 yuan and a score of 5?	餐馆: {人均消费: 100-150元, 评分: 5分, 名称: “”} Restaurant: {per capita consumption: 100-150 yuan, score: 5 points, name: “”}
System	全聚德烤鸭(吕营店)和咸亨酒店(和平店)都是比较火的餐馆，你感兴趣不? Quanjude Roast Duck (Luying) and Xianheng Hotel (Heping) are hot restaurants. Are you interested?	餐馆: {人均消费: 100-150元, 评分: 5分, selectedResults: [全聚德烤鸭(吕营店),咸亨酒店(和平店)] } Restaurant: {per capita consumption: 100-150 yuan, score: 5 points, selectedResults: [Quanjude Roast Duck (Luying), Xianheng Hotel (Heping Branch)] }
User	我选择去全聚德烤鸭(吕营店)吃饭，它家都有什么推荐菜啊? I choose to eat at Quanjude Roast Duck (Luying). What recommendations does it have?	餐馆: { 名称: 全聚德烤鸭(吕营店), 推荐菜: “”, 人均消费:100-150元, 评分: 5分} Restaurant: { name: Quanjude Roast Duck (Luying), recommended dishes: “””, per capita consumption: 100-150 yuan, score: 5 points }
System	推荐菜有清蒸鲈鱼, 盐水鸭干。 Recommended dishes are steamed sea bass and salted duck.	餐馆: { 名称: 全聚德烤鸭(吕营店), 推荐菜: ””, 人均消费: 100-150元, 评分: 5分, selectedResults: [全聚德烤鸭(吕营店)] } Restaurant: { name: Quanjude Roast Duck (Luying), recommended dishes: ””, per capita consumption: 100-150 yuan, score: 5 points, selectedResults: [Quanjude Roast Duck (Luying)] }
User	好的，那它家周边有景点吗? OK, are there any attractions around it?	餐馆: { 名称: 全聚德烤鸭(吕营店), 推荐菜:[清蒸鲈鱼, 盐水鸭干], 人均消费: 100-150元, 评分: 5分, 周边景点: [] } Restaurant: { name: Quanjude Roast Duck (Luying), recommended dishes: [steamed sea bass, dried salted duck], per capita consumption: 100-150 yuan, score: 5 points, attractions nearby: [] }
System	是的，周边景点有北京欢乐谷, 潘家园。 Yes, the surrounding attractions are Beijing Happy Valley, Panjiayuan.	餐馆: { 名称: 全聚德烤鸭(吕营店), 推荐菜: ””, 人均消费: 100-150元, 评分: 5分, 餐馆-周边景点: ””, selectedResults: [全聚德烤鸭(吕营店)] } Restaurant: { name: Quanjude Roast Duck (Luying), recommended dishes: ””, per capita consumption: 100-150 yuan, score: 5 points, attractions nearby: ””, selectedResults: [Quanjude Roast Duck (Luying)] }
User	不错啊，我正打算在餐馆周边找一个游玩时长是1小时- 2小时，票价是20元以下的景点游玩呢，知道哪个比较适合我吗? Yes, I am planning to find a play area around the restaurant that lasts 1 hour-2 hours, and the fare is less than 20 yuan. Do you know which one is more suitable for me?	餐馆: { 名称: 全聚德烤鸭(吕营店), 推荐菜:[清蒸鲈鱼, 盐水鸭干], 人均消费: 100-150元, 评分: 5分, 周边景点: [北京欢乐谷,潘家园] } 景点: { 名称: 出现在id=1的周边景点里, 游玩时间: 1小时- 2小时, 门票: 20元以下} Restaurant: { name: Quanjude Roast Duck (Luying), recommended dishes: [steamed sea bass, dried salted duck], per capita consumption: 100-150 yuan, score: 5 points, attractions nearby: [beijing happy valley, panjiayuan] } Attraction: { name: Appears in surrounding attractions with id = 1 , play time: 1 hour-2 hours, tickets: less than 20 yuan }

System	潘家园是你唯一的选择。
	Panjiayuan is your only choice.
		餐馆: { } //same as above
		景点: { 门票: 20元以下, 游玩时间: 1小时- 2小时, 周边餐馆: 全聚德烤鸭(吕营店), selectedResults: 潘家园}
		Restaurant: { } //same as above
		Attraction: { tickets: less than 20 yuan, play time: 1 hour-2 hours, nearby restaurants: Quanjude Roast Duck (Luying), selectedResults: Panjiayuan }

Table 2: An example of partial dialog for MultiWOZ 2.1 dataset.

Speaker	Utterance	Dialog States
User	I am leaving Cambridge . I need a train that leaves after 13:45 . 我要离开剑桥。我需要一列在13:45之后离开的火车。	
System	On what day do you wish to travel ? 您想在哪一天旅行?	train: {leaveAt: 13:45, departure: cambridge } 火车: {出发时间: 13:45, 出发地: 剑桥}
User	I am traveling on Wednesday and need to go to Birmingham New Street please . 我打算星期三旅行，我需要去伯明翰新街。	
System	TR1766 . It is going to take 163 minutes . can i book it for you ? 您可以乘坐TR1766。这将需要163分钟。需要我为您预订吗?	train: { destination: birmingham new street, leaveAt: 13:45, departure: cambridge, day: wednesday } 火车: {目的地: 伯明翰新街, 出发时间: 13:45, 出发地: 剑桥,天: 星期三}
User	Yes , can you book 4 tickets for me ? 好的，你能帮我预订四张票吗?	
System	I have you booked on TR1766 leaving Cambridge at 18:01 arriving Birmingham New Street at 20:44 . Your reference number is 6L0MANZV . Can I help with anything else ? 我帮您订了TR1766的票，它于18:01离开剑桥，于20:44到达伯明翰新街。您的参考号码是6L0MANZV。您还需要别的帮助吗?	train: { destination: birmingham new street, leaveAt: 13:45, departure: cambridge, day: wednesday, book: {people: 4, booked: [trainID: TR1766, reference: 6L0MANZV] } } 火车: {目的地: 伯明翰新街, 出发时间: 13:45, 出发地: 剑桥, 天: 星期三, 预订: {人数: 4, 预订成功: [火车号: TR1766, 参考号: 6L0MANZV] } },
User	Also looking for a place that has British food and moderately priced . 我也正在寻找一个有英国美食且价格适中的地方。	
System	I have many options for you . They are all in the city center . Do you have any other preferences ? 我有很多选择给你。它们都在市中心。您还有其他偏好吗?	train: { destination: birmingham new street, leaveAt: 13:45, departure: cambridge, day: wednesday, book: {people: 4, booked: [trainID: TR1766, reference: 6L0MANZV] } } 火车: {目的地: 伯明翰新街,出发时间: 13:45, 出发地: 剑桥, 天: 星期三, 预订: {人数: 4, 预订成功: [火车号: TR1766,参考号: 6L0MANZV]} }