Multi-Domain Task-Completion Dialog Challenge

DSTC8 Track Proposal

1 Motivation

This challenge proposal intends to foster progress in two important aspects of dialog systems: dialog complexity and scaling to new domains. First, there is an increasing interest in building complex bots that span over multiple sub-domains to accomplish a complex user goal such as travel planning which may include hotel, restaurant, attraction and so on [1–3]. To advance state-of-the-art technologies for handling complex dialogs, we offer a timely task focusing on multi-domain end-to-end task completion dialog. Second, neural dialog systems require very large datasets to learn to output consistent and grammatically-correct sentences [4–6]. This makes it extremely hard to scale out the system to new domains with limited in-domain data. With this challenge, our goal is to investigate whether sample complexity can decrease with time, *i.e.*, if a dialog system that was trained on a large corpus can learn to converse about a new domain given a much smaller in-domain corpus.

2 Task Description

This track consists of two sub-tasks:

- Participants will build an **end-to-end multi-domain** dialog system for tourist information desk settings.
- Participants will develop **fast adaptation** methods for building a conversation model that generates appropriate domain-specific user responses to an incomplete dialog history.

2.1 End-to-end Multi-domain Task

Previous challenges for dialog systems have greatly helped the research community identify important tasks and rigorously evaluate various approaches. Most prior works tend to focus on individual components in a dialog system, e.g. natural language understanding, dialog state tracking and dialog policy, instead of evaluating the whole system in an end-to-end fashion [7–10]. However, performance improvement of individual components doesn't necessarily translate to that of the entire system. Also, recently, researchers have raced to create end-to-end approaches to minimize laborious hand-coding and error propagation down the pipeline, but there is a scarcity of work comparing such systems with conventional approaches.

In response to such concerns, this challenge task offers various learning resources to allow participants to build end-to-end dialog systems with widely different approaches, ranging from monolithic neural networks to pipelined architectures, and evaluate such systems in an end-to-end fashion. Figure 1a presents a pipelined dialog system as an example and participants are free, and encouraged, to plug in any modules, as long as their systems can complete a predefined task via multi-turn conversations with natural language input and output. In every turn of a conversation, the system needs to understand natural language input generated by the user or the simulator, track dialog states during the conversation, interact with a task-specific dataset, and generate a system response.[1] There are no constraints regarding system architecture and participants are encouraged to explore various approaches such as a monolithic end-to-end neural network shown in Figure 1b or any type of architecture in-between.

[1] The system response must be natural language utterances. But participants can opt to generate dialog-acts when training with a user simulator.

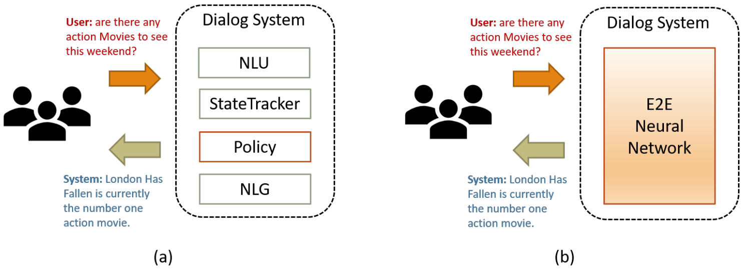
User: are there any action Movies to see this weekend?
Dialog System
NLU
StateTracker
Policy
NLG
System: London Has Fallen is currently the number one action movie.
(a)
User: are there any action Movies to see this weekend?
Dialog System
E2E Neural Network
System: London Has Fallen is currently the number one action movie.
(b)

Figure 1: Illustration of end-to-end task-completion dialog systems.

Specifically, participants are to build a bot for tourist information desk settings based on the recently released MultiWOZ [3] dataset which we enrich with further annotation to support a wider range of learning approaches.

2.1.1 Resources

For this task, the following resources will become available:

- Fully annotated datasets with which participants can train and validate individual components (e.g. NLU, NLG, dialog state tracker, dialog policy) in a dialog system with (un)supervised learning approaches. An example dialog segment can be found in Appendix.
- Backend services to which API calls can be made.
- End-to-end user simulators with NLU and NLG equipped.
- State-of-the art models for NLU, NLG, dialog state tracker, reinforcement learning-based policies any of which participants freely opt to use as part of their system.

2.1.2 Evaluation

To get the best of all different evaluation approaches, this task offers multi-layered evaluation as follows:

- **Corpus-based evaluation:** slot state accuracy, joint state accuracy, BLEU, entropy [2]
- **Simulation-based evaluation:** Task success rate, dialog length, average rewards
- **Crowdworker-based evaluation:** Task success rate, dialog length, irrelevant turn rate, redundant turn rate, user satisfaction score

2.2 Fast Adaptation Task

An end-to-end dialog system trained on many Reddit threads knows how to output grammatically correct, ideally on-topic responses, but will likely fail to predict responses of customers using a goal-oriented travel agency bot. Conversely, a system trained solely on a small dataset of travel agency dialog will fail to produce coherent responses or overfit on the training dialogs.

Participants will develop fast adaptation methods for building a conversation model that generates appropriate domain-specific user responses to an incomplete dialog history. There are three datasets we provide to learn to transfer conversational skills from large-scale open domain to a specific domain, given a small set of in-domain dialogs. We suggest to use techniques such as fine-tuning or learning-to-learn (‘meta-learning’, e.g. [11–13]) to achieve this goal. In appendix A.1, we briefly discusses some approaches.

[2] Corpus-based evaluation applies only when the system outputs labels or natural language responses for a dialog context taken from the test corpus

2.2.1 Training Resources

Reddit Dataset We constructed a corpus of dialogs from Reddit submissions and comments spanning November 2017 through October 2018. Content is selected from a curated list of one thousand high-traffic subreddits. Our extraction and filtering methodology is based on that used in the DSTC7 sentence generation task [14], the key difference being we sample at most two threads per submission. The corpus consists of five million training dialogs, with an additional one million dialogs reserved for validation. We provide pre-processing code for Reddit data so that all participants work on the same corpus.

Goal-Oriented Corpus MetaLWOz We collected 37 884 goal-oriented dialogs via crowd-sourcing using a *Wizard of Oz*, or *WOz* scheme. These dialogs span 47 domains and 227 tasks and are particularly suited for meta-learning dialog models. For each dialog, we paired two crowd-workers, one had the role of being a bot, and the other one was the user. Both were given a domain and a task. Examples of domains are: bus schedule, apartment search, alarm setting, banking and event reservation. We defined several tasks per domain. An example of task for the bus schedule domain is: *Inform the user that the bus stop they are asking about has been moved two blocks north* on the bot side, and *Ask if a certain bus stop is currently operational* on the user side. Statistics of the dataset are provided in Table 2. We will release an analysis of the complexity of the domains to provide some insights on the adaptation challenges. Note that all entities were invented by the crowd-workers (for instance, the address of the bus stop) and the goal of this challenge is to automate the user utterances and not the bot utterances. Samples from this dataset are given in the Appendix.

Baseline Model We will provide code for a baseline meta-learner that participants can build upon. We will also release evaluation results for this model at the beginning of the challenge.

2.2.2 Evaluation Resources

We will evaluate models on subsets of the MultiWoz [3] dataset, where dialogs are limited to a single domain. We will provide both the splits of the MultiWoz dataset and the NLU from Section 2.1.1 for automatic evaluation.

2.2.3 Evaluation

We will evaluate responses by the domain-adapted dialog model using two metrics:

- **Automatic metrics:** A small set of MultiWoz dialogs is provided to the model, which is then asked to respond to an incomplete dialog. Intents and slot values correctly detected by the baseline NLU in the response serve as an indicator that the domain adaptation was successful.
- **Human evaluation:** Human annotators will be asked to judge the appropriateness, informativeness and utility of the responses [15] *given the domain.*

3 Organizers

Adam Atkinson*, Hannes Schulz*, Jianfeng Gao†, Kaheer Suleman*, Layla El Asri*, Mahmoud Adada*, Minlie Huang‡, Shikhar Sharma*, Sungjin Lee†, Wendy Tay* and Xiujun Li†. [3]

Contact information:

Sungjin Lee (`sule@microsoft.com`) and Hannes Schulz (`hannes.schulz@microsoft.com`).

[3] Microsoft Research in Montréal*, Microsoft Research AI† and Tsinghua University‡

References

[1] B. Peng, X. Li, L. Li, J. Gao, A. Çelikyilmaz, S. Lee, and K. Wong. “Composite Task-Completion Dialogue Policy Learning via Hierarchical Deep Reinforcement Learning”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 2017.

[2] L. E. Asri, H. Schulz, S. Sharma, J. Zumer, J. Harris, E. Fine, R. Mehrotra, and K. Suleman. “Frames: A corpus for adding memory to goal-oriented dialogue systems”. In: *arXiv preprint arXiv:1704.00057* (2017).

[3] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić. “MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 2018.

[4] O. Vinyals and Q. V. Le. “A Neural Conversational Model”. In: *arXiv:1506.05869* (2015).

[5] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky. “Deep Reinforcement Learning for Dialogue Generation”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 2016.

[6] T.-H. Wen, Y. Miao, P. Blunsom, and S. Young. “Latent Intention Dialogue Models”. In: *Proceedings of the International Conference on Machine Learning.* 2017.

[7] J. Williams, A. Raux, D. Ramachandran, and A. Black. “The dialog state tracking challenge”. In: *Proceedings of the SIGDIAL 2013 Conference.* 2013, pp. 404–413.

[8] M. Henderson, B. Thomson, and J. D. Williams. “The second dialog state tracking challenge”. In: *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL).* 2014, pp. 263–272.

[9] M. Henderson, B. Thomson, and J. D. Williams. “The third dialog state tracking challenge”. In: *Spoken Language Technology Workshop (SLT), 2014 IEEE.* IEEE. 2014, pp. 324–329.

[10] S. Kim, L. F. D’Haro, R. E. Banchs, J. D. Williams, and M. Henderson. “The fourth dialog state tracking challenge”. In: *Dialogues with Social Robots.* Springer, 2017, pp. 435–449.

[11] C. Finn, P. Abbeel, and S. Levine. “Model-agnostic meta-learning for fast adaptation of deep networks”. In: *Proceedings of the International Conference on Machine Learning.* 2017.

[12] S. Ravi and H. Larochelle. “Optimization as a model for few-shot learning”. In: *Proceedings of the International Conference on Learning Representations.* 2017.

[13] J. Gu, Y. Wang, Y. Chen, K. Cho, and V. O. Li. “Meta-Learning for Low-Resource Neural Machine Translation”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 2018.

[14] K. Yoshino, C. Hori, J. Perez, L. F. D’Haro, L. Polymenakos, C. Gunasekara, W. S. Lasecki, J. Kummerfeld, M. Galley, C. Brockett, J. Gao, B. Dolan, S. Gao, T. K. Marks, D. Parikh, and D. Batra. “The 7th dialog system technology challenge”. In: *arXiv preprint* (2018).

[15] M. Galley, C. Brockett, X. Gao, B. Dolan, and J. Gao. “End-to-End conversation Modeling: DSTC7 Task 2 Description”. In: *DSTC7 workshop (forthcoming).*

[16] T. Munkhdalai, X. Yuan, S. Mehri, and A. Trischler. “Rapid adaptation with conditionally shifted neurons”. In: *Proceedings of the International Conference on Machine Learning*. 2018.

[17] A. Rastogi, D. Hakkani-Tür, and L. P. Heck. “Scalable Multi-Domain Dialogue State Tracking”. In: *Proceedings of Automatic Speech Recognition and Understanding*. 2017.

[18] I. Shalyminov, A. Eshghi, and O. Lemon. “Multi-Task Learning for Domain-General Spoken Disfluency Detection in Dialogue Systems”. In: *Proceedings of the Workshop Series on the Semantics and Pragmatics of Dialogue*. 2018.

Appendix

Speaker	Utterance	Annotation
User	I am looking for a place to to stay that has cheap price range it should be in a type of hotel	**Dialog acts** {”Hotel-Inform”: [[”Pricerange”, ”cheap”]]}
System	Okay, do you have a specific area you want to stay in?	**State** {“hotel”: { ”name”: ”not mentioned”, ”area”: ”not mentioned”, ”parking”: ”not mentioned”, ”pricerange”: ”cheap”, ”stars”: ”not mentioned”, ”internet”: ”not mentioned”, ”type”: ”hotel”}}... **Dialog acts** {”Hotel-Request”: [[”Area”, ”?”]]}
User	no, i just need to make sure it’s cheap. oh, and i need parking.	**Dialog acts** {“negate”, ”Hotel-Inform”: [[”Pricerange”, ”cheap”], [“Parking”, “yes”]]}
System	I found 1 cheap hotel for you that includes parking. Do you like me to book it?	**State** {“hotel”: { ”name”: ”not mentioned”, ”area”: ”not mentioned”, ”parking”: ”yes”, ”pricerange”: ”cheap”, ”stars”: ”not mentioned”, ”internet”: ”not mentioned”, ”type”: ”hotel”}}... **Dialog acts** {”Hotel-Inform”: [[”Price”, ”cheap”], [”Choice”, ”1”]], [”Parking”, ”none”]] }

Table 1: An incomplete example dialog for the multi-domain dialog task.

	Mean	Minimum	Maximum
Number of tasks per domain	4.8	3	11
Number of dialogs per domain	806.0	288	1990
Number of dialogs per task	167.6	32	285
Number of turns per dialog	11.4	10	46

Table 2: Statistics of the goal-oriented dataset

A Meta-Learning

Recent advances in meta-learning have allowed to train neural networks to adapt to a new task after seeing only a few examples of the task. Most progress on the topic has been made in computer vision [11, 12] but results on some natural language processing tasks are encouraging [13, 16]. This proposal aims to leverage these new techniques for neural multi-domain dialog systems [1, 3, 17, 18].

A.1 Possible Approaches for Challenge Participants

Fine-Tuning Train an end-to-end dialog base model on reddit dialogs, use the provided sample of in-domain dialogs (MetaLWOz or MultiWoz) to fine-tune it with a cross-validated learning rate.

Learn how to fine-tune use the technique outlined by Ravi and Larochelle [12] to *learn* how to finetune the base model given the gradients on MetaLWOz or MultiWoz.

MAML use the technique outlined by Finn, Abbeel, and Levine [11] or Gu et al. [13] to learn a base model that is easy to adapt. For example, one might use samples from Reddit to compute a first order gradient and samples from MetaLWOz to compute a second order gradient. Finally, the model could be adapted using dialog samples from MultiWoz for evaluation.

B The MetaLWOz Dataset

B.1 List of Domains

Calendar update, pizza ordering, movie listings, event reservation, weather checking, flight booking, contact information update, restaurant reservation, playlist editing, looking up information, shopping, store details, sports information, quote of the day, how to, prompt generation, library request, banking, restaurant picking, name suggestion, vacation ideas, city information, music suggestion, agreement, pet advice, apartment search, Guinness record check, geography, alarm setting, contact management, phone settings, appointment reminder, home, policy, decision making, catalogue lookup, ski, bus schedule, insurance, information retrieval, sorting, scam lookup, time zone, play schedule, game rules, wedding planning, status check, gift ideas, tourism, hotel reservation and phone plan.

B.2 Samples of Dialogs

Domain: Ski
User Task: You want to know if there are good ski hills an hour's drive from your current location.
Bot Task: Tell the user that there are no ski hills in their immediate location.
Bot Hello how may I help you?
User Is there any good ski hills an hour's drive from my current location?
Bot I'm sorry to inform you that there are no ski hills in your immediate location
User Can you help me find the nearest?
Bot Absolutely! It looks like you're about 3 hours away from Bear Mountain. That seems to be the closest.
User Hmm.. sounds good
Bot Alright! I can help you get your lift tickets now!When will you be going?
User Awesome! please get me a ticket for 10pax
Bot You've got it. Anything else I can help you with?
User None. Thanks again!

Bot No problem!
Example of input/output for this dialog:
Input: domain = Ski, user task = You want to know if there are good ski hills an hour's drive from your current location., bot task = Tell the user that there are no ski hills in their immediate location, dialog history = Hello how may I help you?
Output: user response = Is there any good ski hills an hour's drive from my current location?

Domain: Game rules
User Task: Ask how to win at the text adventure Zork. If the Bot cannot help you, ask them what games it can help you with.
Bot Task: Tell the user that you are programmed to help them play games, not win them. If the user starts talking about other games, go along with them and try to help.
Bot Hello how may I help you?
User I want to know how I can win playing Zork?
Bot I am programmed to help you play games, not win them
User What games can you help me with?
Bot Any game
User Can you specify a certain game?
Bot Zork
User What are the rules of Zork?
Bot You have to follow through a cave adventure
User What is the ending of the game?
Bot The user comes out of the cave
Example of input/output for this dialog:
Input: domain = game rules, user task = Ask how to win at the text adventure Zork. If the Bot cannot help you, ask them if what games it can help you with., bot task = Tell the user that you are programmed to help them play games, not win them. If the user starts talking about other games, go along with them and try to help., dialog history = Hello how may I help you?; I want to know how I can win playing Zork?; I am programmed to help you play games, not win them
Output: user response = What games can you help me with?

Domain: Music suggestion
User Task: Tell the bot that you like a certain genre of music, but don't prompt it to give you suggestions
Bot Task: Provide suggestions, even if the user doesn't ask for them (or tells you they don't need any)
Bot Hello how may I help you?
User Hello. I like classic rock music.
Bot Dream on Aerosmith
User No, I'm not a huge fan of that song. Is there anything else available?
Bot Sweet Home Alabama
User Well, I do like Lynyrd Skynyrd, but that song is so cliche. Do you have anything else by the same band?
Bot Free bird
User Yes, that is the perfect song for how I am feeling right now!
Bot Great, Is there anything else?
User No, but I really appreciate your assistance. Thank you so much!

Bot Youre welcome
Example of input/output for this dialog:
Input: domain = music suggestion, user task = Tell the bot that you like a certain genre of music, but don't prompt it to give you suggestions, bot task = Provide suggestions, even if the user doesn't ask for them (or tells you they don't need any)., dialog history = Hello how may I help you?; Hello. I like classic rock music.; Dream on Aerosmith; No, I'm not a huge fan of that song. Is there anything else available?; Sweet Home Alabama;
Output: user response = Well, I do like Lynyrd Skynyrd, but that song is so cliche. Do you have anything else by the same band?

Domain: Information retrieval
User Task: Describe a movie where these guys have to climb a mountain and then throw a thing into the mountain and ask the bot if they know what the movie is called.
Bot Task: Ask the user to elaborate as much as possible about the movie they are asking about, and then suggest they're thinking of Lord of the Rings.
Bot Hello how may I help you?
User I'm trying to remember this movie. Can you help?
Bot I would be happy to help. Can you tell me what it is about?
User In it these guys have to climb up a mountain. Then they throw something into the mountain.
Bot Do you know what genre the movie is, such as comedy, horror, sci-fi, or fantasy?
User Not sure about that.
Bot Do you any of the actor or character names in the movie?
User No... All I can remember is the stuff about the mountain.
Bot Is it a new movie, or an older movie?
User I don't know.
Bot It sounds like the movie you are talking about could be Lord of the Rings. Does that sound right?
User It could be that. Thanks for the help.
Example of input/output for this dialog:
Input: domain = information retrieval, user task = Describe a movie where these guys have to climb a mountain and then throw a thing into the mountain and ask the bot if they know what the movie is called., bot task = Ask the user to elaborate as much as possible about the movie they are asking about, and then suggest they're thinking of Lord of the Rings., dialog history = Hello how may I help you?; I'm trying to remember this movie. Can you help?; I would be happy to help. Can you tell me what it is about?; I would be happy to help. Can you tell me what it is about?
Output: user response = In it these guys have to climb up a mountain. Then they throw something into the mountain.