End-to-end Task Completion Dialog Challenge

Description

This challenge is aiming to advance state-of-the-art technologies for building end-to-end task-completion dialog systems and offer multi-layered evaluation to understand discrepancies among different evaluation methods and yield more robust results. This track consists of two sub-tasks:

1) building a dialog system for the *Movie* domain;
2) building a multi-domain dialog system for *tourist information desk* settings.

Figure 1a presents a pipelined dialogue 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 dialogue 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.

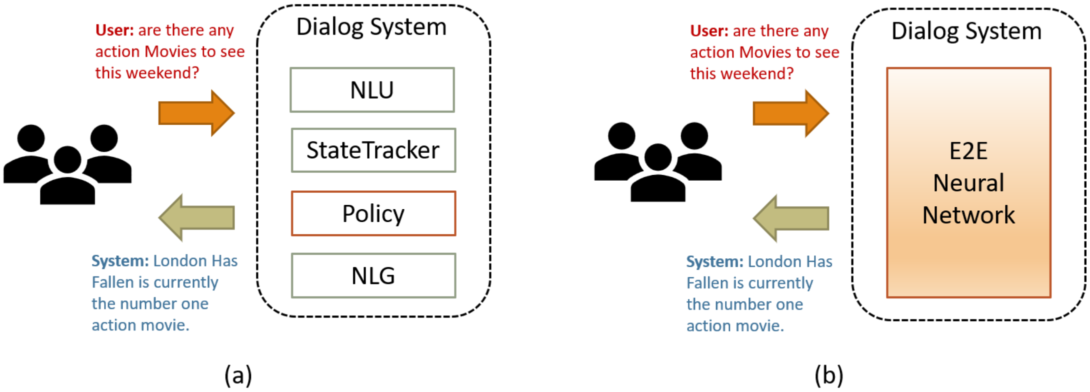
User: are there any action Movies to see this weekend?
Dialog System
NLU
StateTracker
Policy
NLG
System: London Has Fallen is currently the number one action movie.
(a)
User: are there any action Movies to see this weekend?
Dialog System
E2E Neural Network
System: London Has Fallen is currently the number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

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

Previous challenges for dialog systems have greatly helped the research community identify important tasks and rigorously evaluate various approaches. However, such prior work tends 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 [1-4]. 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 along the pipeline, but there is a scarce of work comparing such systems with conventional approaches. In response to such concerns, this challenge offers various learning resources to allow participants to build end-to-end dialog systems with widely different approaches,

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

ranging from monolithic neural networks to pipelined architectures, and evaluate such systems in an end-to-end fashion.

On the other hand, development of more complex and advanced dialog systems has also introduced evaluation challenges and effort has been directed toward different evaluation settings such as corpus-based automatic evaluation, simulation-based evaluation and crowdworker-based evaluation. Despite its simplicity and popularity, it is well known that corpus-based evaluation is limited due to some critical issues such as exposure bias and covariate shift problems. Another line of research has dealt with automatic evaluation with user simulations to alleviate the data intensiveness problem of reinforcement learning approaches for policy optimization and enable quick assessment with different user models. However, simulated users do not fully reflect the natural conditions in which the system would be used in interactions with human users. To move toward human evaluation, crowdworkers have been playing increasingly important roles in making human users readily accessible. However, prior observations have found significant differences between paid, informed users and unpaid, real users [5-7]. For evaluation with paid, informed users, task descriptions and user requirements may be unrepresentative of some situations that occur in authentic usage contexts. A more serious problem is the fact that the subjects adapt themselves to the given scenario. Although real situations are generally regarded as providing the best conditions for evaluation, evaluation in real situations is costly due to the complexity of the evaluation setup. Considering pros and cons of different evaluation approaches, this challenge offers multi-layered evaluation consisting of corpus and simulated user-based automatic evaluation and human evaluation with crowdworkers. Furthermore, for the Movie domain, pre-screened top-quality systems will get connected to Microsoft Bing's movie finding service for real user evaluation.

Lastly, there is increasing interest in building complex bots that span over multiple sub-domains to accomplish a complex user goal such as tourist information provision which may include hotel, restaurant, attraction and so on [8-10]. To foster further research, this challenge offers a timely sub-task focusing on multi-domain end-to-end task completion dialog. Specifically, participants are to build a bot for tourist information desk settings based on the recently released MultiWOZ [10] dataset which we enrich with further annotation to support a wider range of learning approaches.

Challenge resources

This challenge aims toward the building of shared infrastructure, corpora, and open-source system components that allow rapid prototyping and development of systems. For each sub-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.

Evaluation methods

To get the best of all different evaluation approaches, this challenge 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
- Real user-based evaluation: Task success rate, dialog length, irrelevant turn rate, redundant turn rate, user satisfaction score[3]

Organizers

Sungjin Lee, Microsoft Research, sule@microsoft.com

Xiujun Li, Microsoft Research, xiul@microsoft.com

Minlie Huang, Tsinghua University, aihuang@mail.tsinghua.edu.cn

Jianfeng Gao, Microsoft Research, jfgao@microsoft.com

References

[1] Williams, Jason, et al. "The dialog state tracking challenge." *Proceedings of the SIGDIAL 2013 Conference*. 2013.

[2] Henderson, Matthew, Blaise Thomson, and Jason D. Williams. "The second dialog state tracking challenge." *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 2014.

[3] Henderson, Matthew, Blaise Thomson, and Jason D. Williams. "The third dialog state tracking challenge." *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014.

[4] Kim, Seokhwan, et al. "The fourth dialog state tracking challenge." *Dialogues with Social Robots*. Springer, Singapore, 2017. 435-449.

[5] Rudnicky, A., C. Bennett, A. Black, A. Chotimongkol, K. Lenzo, A. Oh, and R. Singh (2000). Task and Domain Specific Modelling in the Carnegie Mellon Communicator system. In Proceedings of ICSLP2000, Beijing, China, vol. II, pp. 130–133.

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

[3] Real user-based evaluation is only available for sub-task 1.

[6] Ai, H., Raux, A., Bohus, D., Eskenazi, M., and Litman, D. (2007). Comparing spoken dialog corpora collected with recruited subjects versus real users. Proceedings of the 8th SIGDial Workshop on Discourse and Dialogue, Antwerp, Belgium.

[7] Black, Alan W., et al. "Spoken dialog challenge 2010: Comparison of live and control test results." *Proceedings of the SIGDIAL 2011 Conference*. Association for Computational Linguistics, 2011.

[8] Peng, Baolin, et al. "Composite Task-Completion Dialogue Policy Learning via Hierarchical Deep Reinforcement Learning." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

[9] Asri, Layla El, et al. "Frames: A corpus for adding memory to goal-oriented dialogue systems." *arXiv preprint arXiv:1704.00057* (2017).

[10] Budzianowski, Paweł, et al. "MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018.

Appendix

Speaker	Utterance	Annotation
User	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"]]}