A Challenge on Semi-Supervised and Reinforced Task-Oriented Dialog Systems

Zhijian Ou[1], Junlan Feng[2], Juanzi Li[1],

[1] Tsinghua University, Beijing, China
[2] China Mobile Research Institute, Beijing, China
Version 1 (April 19, 2022)

Overview

Task-oriented dialogue (TOD) systems are designed to assist users to accomplish their goals, and have gained more and more attention recently in both academia and industry with the current advances in neural approaches [1]. A TOD system typically consists of several modules, which track user goals to update dialog states, query a task-related knowledge base (KB) using the dialog states, decide actions and generate responses. Unfortunately, building TOD systems remains a label-intensive, time-consuming task for two main reasons. First, training neural TOD systems requires manually labeled dialog states and system acts (if used), in both traditional modular approach [12, 8] and recent end-to-end trainable approach [11, 7, 4, 9, 13]. Second, it is often assumed that a task-related knowledge base is available. But for system development from scratch in many real-world tasks, expert labors are needed to construct the KB from annotating unstructured data. Thus, the labeled-data scarcity challenge hinders efficient development of TOD systems at scale.

Remarkably, unlabeled data are often easily available in many forms such as human-to-human dialogs, open-domain text corpus, and unstructured knowledge documents. This has motivated the development of semi-supervised learning (SSL) [15], which aims to leverage both labeled and unlabeled data, for both information extraction to construct the knowledge base and building the TOD system itself. Additionally, although it has long been recognized that TOD systems could be formulated as Markov Decision Processes (MDPs) and trained via reinforcement learning (RL) for the policy learning for the agent [12], it remains very challenging to build reinforced TOD systems due to large language action spaces. There are significant individual research threads, including semi-supervised information extraction [6, 10], using pre-trained language models [2, 5] or latent variable models [14] for semi-supervised TOD systems, grounded response generation with unstructured knowledge sources [3], reinforcement training of the system from interactions with user simulators, and so on.

The purpose of this challenge is to invite researchers from both academia and industry to share their perspectives on building semi-supervised and reinforced TOD systems and to advance the field in joint effort. A shared task is organized for benchmarking and stimulating relevant researches, with a newly released large-scale, multi-domain TOD dataset which consists of 100,000 real-world dialogs.

Techniques of interest

This challenge encourages submissions on building semi-supervised and reinforced TOD systems. All types of semi-supervised techniques are welcome, such as, to name a few, pre-training, self-training, self-supervised, weakly-supervised, transfer learning for zero-shot or few-shots, latent-variable modeling, and domain adaptation, and data augmentation. Both online and offline RL techniques are welcome.

Possible techniques include, but are not limited to, the following:

- General techniques for task-oriented dialog systems
- Semi-supervised information extraction and knowledge modeling
- Grounded dialog with unstructured knowledge sources
- Semi-supervised task-oriented dialog systems
- Reinforced task-oriented dialog systems
- User simulators

Shared task

We introduce a new shared task, aiming to benchmark semi-supervised and reinforced task-oriented dialog systems, built for automated customer-service for mobile operators. The task consists of two tracks: information extraction from dialog transcripts (Track 1) and task-oriented dialog systems (Track 2).

An important feature for this shared task is that we release around 100,000 dialogs (in Chinese), which come from real-world dialog transcripts between real users and customer-service staffs from China Mobile, with privacy information anonymized. We call this dataset as MCSD (mobile customer-service dialog) dataset, which differs from existing TOD datasets in both size and nature significantly. To the best of our knowledge, MCSD is

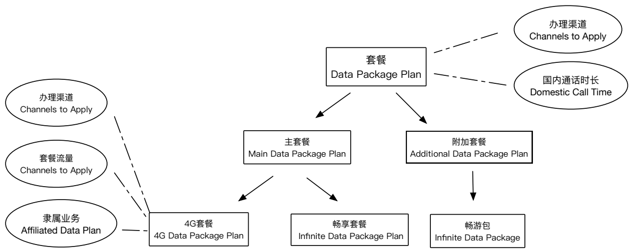

Figure 1: An example of a schema for the "data package plan" domain, with concepts (in rectangles) and attributes (in circles).

not only the largest publicly available multi-domain TOD dataset, but also consists of real-life data (namely collected in real-world scenarios). For comparison, the widely used MultiWOZ dataset consists of 10,000 dialogs and is in fact simulated data (namely collected in a Wizard-of-Oz simulated game).

We provide a schema, and only 10,000 dialogs are labeled by crowdsourcing, while the remaining 90,000 dialogs are unlabeled. The teams are required to use this mix of labeled and unlabeled data to train information extraction models (Track 1), which could provide a knowledge base for Track 2, and train TOD systems (Track 2), which could work as customer-service bots. We put aside 5,000 dialogs as evaluation data.

Track 1: Information Extraction from dialog transcripts

Motivation In a task-oriented dialog system, after dialog state tracking, the system needs to query a task-related knowledge base. The query result is important for system action decision. For system development from scratch in many real-world tasks, the knowledge base is often not readily available for training TOD systems. Traditionally, expert labors are needed to construct the knowledge base.

Given a mix of labeled and unlabeled dialog transcripts, Track 1 examines the task of training information extraction models to construct the "local" knowledge base for each dialog, which will be needed in training TOD systems in Track 2. The knowledge base is local in the sense that the mentioned entities with their mentioned attributes are extracted across all turns in a dialog, but there is no information fusion between dialogs[1]. With such knowledge base, we will be able to drive the training of the TOD system.

Schema A schema is a collection of hierarchical concepts with attributes, used to organize and interpret information in a domain. We will provide a manually designed schema including concepts and their attributes, which is illustrated in Figure 1.

[1]We leave information fusion across dialogs for future study.

Task definition Based on the schema, we define two sub-tasks.

1) Entity extraction. This sub-task is to extract entities with their corresponding concepts, which are mentioned in a dialog session. In real-life dialogs, an entity may be mentioned in different surface forms, which need to be extracted. For example, "50元流量包" (50 Yuan data package plan) may have a number of different mentions in a multi-turn dialog: "50元那个业务" (50 Chinese Yuan plan), "那个流量包" (that package plan), "刚才那个业务" (that plan). Thus, entity extraction for the MCSD dataset is more challenging than classic named entity recognition tasks (e.g., extracting person names), due to the informal, verbalized and loose form of the customer-service dialogs.

2) Slot filling. This sub-task is to extract slot values for entity slots (i.e., attributes). It is formulated as a sequence labeling task for the pre-defined slots in the schema. For example, in sentence "10GB套餐业务每月的费用是50块钱。" (The price for 10GB data package plan is 50 Chinese Yuan per month), "每月的费用是50块钱" (50 Chinese Yuan per month) will be labeled as plan price slot. An entity may have several mentions in a dialog, and the slots and values for an entity may scatter in multi-turn dialogs. Thus, the task requires entity resolution and assigning slot-value pairs to the corresponding entity. After entity extraction and slot filling, a local knowledge base will be constructed with all extracted entities with their attributes for each dialog.

Evaluation Given a dialog in testing, the trained information extraction model is used to extract entities together with slot values. We will evaluate and rank the submitted models by the extraction performance on test set. The evaluation metrics are Precision, Recall and F1. As for entity extraction, the metrics are at entity level: an entity is extracted correctly if and only if the mention span of the entity is labeled as the corresponding entity type (i.e., concept). As for slot filling, the metrics are at triple level: an entity-slot-value triple is extracted correctly if and only if 1) the mention span of the slot value is labeled as the corresponding slot type. 2) the slot-value pair is correctly assigned to the corresponding entity.

The F1 scores will be the main ranking basis on leaderboard. We will provide the following scripts and tools for the participants: 1) Baseline models for both sub-tasks; 2) Evaluation scripts to calculate the metrics.

Track 2: Task-Oriented Dialog Systems

Motivation Most existing TOD systems require large amounts of annotations of dialog states and dialog acts (if used), which are time-consuming and labor-intensive. Track 2 examines the task of training a TOD system over the mix of labeled and unlabeled dialog transcripts.

Task definition For every labeled dialog, the annotations contain user information (such as user's data package plan, payment records and so on), which is needed for the customer-service agent to complete the whole dialog session. This task has two features about its knowledge base, which are different from those in other TOD tasks:

1. The user information about each individual user is the basic knowledge to complete the dialog, which is refereed to as the local knowledge base in Track 1;
2. No global knowledge base is used.

The teams are encouraged to utilize the unlabeled dialogs provided in the MCSD dataset.

Connection with Track1 For every unlabeled dialog in training, the organizers will provide extracted user information by running the baseline of Track 1, which the teams can use as the knowledge base. The teams are allowed and encouraged to use their own information extraction models, built in Track 1, to construct the knowledge base for training TOD systems in Track 2.

Evaluation In order to measure the performance of TOD systems, the evaluation data are additionally labeled with user goals, in addition to the labeled user information. User goal means the main purpose of the user engaged in a dialogue, according to which user will talk to the system. User goals over the evaluation data are used by the organizers to calculate the metric, and not provided to the teams. Note that we do not need to label user goals for training data. The main metrics are Success rate and BLEU score. Success rate is the percentage of generated dialogs that achieve user goals. BLEU score evaluates the fluency of generated responses. A combined score is computed as BLEU+Success.

The combined scores will be the main ranking basis on leaderboard. We will provide the following scripts and tools for the participants: 1) A baseline system; 2) Evaluation scripts to calculate the metrics.

References

[1] Jianfeng Gao, Michel Galley, and Lihong Li. *Neural approaches to conversational AI: Question answering, task-oriented dialogues and social chatbots*. Now Foundations and Trends, 2019.

[2] Ehsan Hosseini-Asl et al. "A simple language model for task-oriented dialogue". In: *NeurIPS*. 2020.

[3] Seokhwan Kim et al. "Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access". In: *arXiv preprint arXiv:2006.03533* (2020).

[4] Wenqiang Lei et al. "Sequicity: Simplifying Task-oriented Dialogue Systems with Single Sequence-to-Sequence Architectures". In: *ACL*. 2018.

[5] Baolin Peng Chunyuan Li et al. "SOLOIST: Building Task Bots at Scale with Transfer Learning and Machine Teaching". In: *Trans. of the Association for Computational Linguistics (TACL)* (2021).

[6] Chengjiang Li et al. "Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model". In: *EMNLP-IJCNLP*. 2019.

[7] Bing Liu and Ian Lane. "An End-to-End Trainable Neural Network Model with Belief Tracking for Task-Oriented Dialog". In: *Proc. Interspeech 2017* (2017).

[8] Nikola Mrkšić et al. "Neural Belief Tracker: Data-Driven Dialogue State Tracking". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2017.

[9] Lei Shu et al. "Flexibly-Structured Model for Task-Oriented Dialogues". In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. 2019.

[10] Yunfu Song et al. "Upgrading CRFS to JRFS and its Benefits to Sequence Modeling and Labeling". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020.

[11] Tsung-Hsien Wen et al. "A Network-based End-to-End Trainable Task-oriented Dialogue System". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 2017.

[12] Steve Young et al. "POMDP-based statistical spoken dialog systems: A review". In: *Proceedings of the IEEE* (2013).

[13] Yichi Zhang, Zhijian Ou, and Zhou Yu. "Task-Oriented Dialog Systems that Consider Multiple Appropriate Responses under the Same Context". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*. 2020.

[14] Yichi Zhang et al. "A Probabilistic End-To-End Task-Oriented Dialog Model with Latent Belief States towards Semi-Supervised Learning". In: *EMNLP*. 2020.

[15] Xiaojin Zhu. "Semi-supervised learning literature survey". In: *Technical report, University of Wisconsin-Madison* (2006).