ConvLab-2

DSTC9 Track 2: Multi-domain Task-oriented Dialog Challenge II

ConvLab-2

ConvLab-2 is an open-source toolkit that enables researchers to build task-oriented dialog systems with state-of-the-art models, perform an end-to-end evaluation, and diagnose the weakness of systems. As the successor of ConvLab, ConvLab-2 inherits ConvLab's framework but integrates more powerful dialog models and supports more datasets.

ConvLab-2 Code

ConvLab-2 Paper

DSTC9 Track 2: Multi-domain Task-oriented Dialog Challenge II

As part of DSTC9, Microsoft Research and Tsinghua University are hosting Multi-domain Task-oriented Dialog Challenge II, aiming to solve two tasks in the multi-domain task completion setting: i) End-to-end Multi-domain Task Completion Dialog, ii) Cross-lingual Multi-domain Dialog State Tracking.

Dialog Challenge Page

End-to-end Multi-domain Task Completion Dialog Task

Human Evaluation Leaderboard

Rank	Team ID	Best Spec #	Average Success Rate	Success Rate w/ DB Grounding	Success Rate w/o DB Grounding	Language Understanding Score	Response Appropriateness Score	Turns
1	1	Submission5	74.8	70.2	79.4	4.54	4.47	18.5
1	2	Submission1	74.8	68.8	80.8	4.51	4.45	19.4
3	7	Submission4	72.3	62	82.6	4.53	4.41	17.1
4	6	Submission1	70.6	60.8	80.4	4.41	4.41	20.1
5	3	Submission3	67.8	60	75.6	4.56	4.42	21
6	4	Submission2	60.3	51.4	69.2	4.49	4.22	17.7
7	5	Submission2	58.4	50.4	66.4	4.15	4.06	19.7
8	9	Submission1	55.2	43.2	67.2	4.15	3.98	19.2
9	8	Submission1	35	26	44	3.27	3.15	18.5

10	10	Submission4	19.5	6	33	3.23	2.93	18.8
N/A	Baseline	Baseline	69.6	56.8	82.4	4.34	4.18	18.5

(Team 1 & Team 2 cochampion)

- Average Success = (Success Rate w/ DB Grounding + Success Rate w/o DB Grounding) / 2
- Success Rate w/o DB Grounding: Annotation provided by MTurkers (Success or Fail). They have no idea whether the provided values are consistent with DB.
- Success Rate w/ DB Grounding: It is a success if and only if
  - MTurkers mark the dialog as `Success`
  - The provided `request` slot values plus inform slot values can be found in DB.

Automatic Evaluation Leaderboard

Rank	Team ID	Best Spec #	Success Rate	Complete Rate	Book Rate	Inform P/R/F1	Turn(succ/all)
1	1	Submission3	93	95.2	94.6	84.1/96.2/88.1	12.5/12.7
2	2	Submission5	91.4	96.9	96.2	80.2/97.3/86.0	15.3/15.7
3	3	Submission1	90.8	94.4	96.7	81.0/95.4/85.9	13.4/13.6
4	4	Submission2	89.8	94.6	96.3	72.4/96.0/80.1	15.1/15.8
5	5	Submission2	83.3	88.5	89.1	81.1/90.3/83.5	13.5/13.8
6	6	Submission1	67.7	88.5	90.8	70.4/85.6/75.2	12.8/14.2
7	7	Submission4	57.8	87.1	85	68.7/81.6/72.6	13.7/16.4
8	8	Submission1	52.6	66.9	66.7	57.5/80.7/64.8	13.2/22.5
9	9	Submission1	44.4	50	26.5	57.9/64.5/58.9	12.2/14.6
10	10	Submission4	21.4	40.7	0	55.4/60.0/54.1	11.0/25.9
N/A	Baseline	Baseline	85	92.4	91.4	79.3/94.9/84.5	13.8/14.9

Cross-lingual Multi-domain Dialog State Tracking Task

- The evaluation test sets for both MultiWOZ and CrossWOZ are newly collected data.
- The reported numbers are the average of evaluation results for the public test set and private test set.
- The joint accuracy for each test set is also listed.

MultiWOZ Leaderboard

Rank	Team ID	Best Spec #	Joint Accuracy	Slot Accuracy	P/R/F1	Joint Accuracy (public)	Joint Accuracy (private)
1	1	Submission1	62.37	98.09	92.15/94.02/93.07	62.70	62.03
2	2	Submission2	62.08	98.10	90.61/96.20/93.32	63.25	60.91
3	3	Submission3	30.13	94.40	87.07/74.67/80.40	30.53	29.72

CrossWOZ Leaderboard

Rank	Team ID	Best Spec #	Joint Accuracy	Slot Accuracy	P/R/F1	Joint Accuracy (public)	Joint Accuracy (private)
1	3	Submission3	16.86	89.11	68.26/62.85/65.45	16.82	16.89
2	1	Submission1	15.28	90.37	65.94/78.87/71.82	15.19	15.37
3	2	Submission2	13.99	91.92	72.63/78.90/75.64	14.41	13.58

CrossWOZ Leaderboard (Updated Evaluation)

(Candidate lists from selectedResults are used to correct name labels. Please check the competition description page and ConvLab-2 codebase for details.)

Rank	Team ID	Best Spec #	Joint Accuracy	Slot Accuracy	P/R/F1	Joint Accuracy (public)	Joint Accuracy (private)
1	2	Submission2	32.30	94.35	81.39/82.25/81.82	32.70	31.89
2	1	Submission1	23.96	92.94	74.96/83.41/78.96	23.45	24.47
3	3	Submission3	15.31	89.70	74.78/64.06/69.01	14.25	16.37