A Summary of the Dialog State Tracking Challenge Results

Wencan Luo

September 19, 2013

1 Results

9 teams entered the DSTC, submitting a total of 27 trackers.

There are three schedules for determining which turns to include in each evaluation.

Schedule 1: Include all turns.

Schedule 2: Include a turn for a given concept only if that concept either appears on the SLU N-Best list in that turn, or if the systems action references that concept in that turn.

Schedule 3: Include only the turn before the system starts over from the beginning, and the last turn of the dialog.

It evluated in 5 different data sets. The number of turns under different schedules are shown in Table 1

	test1	test2	test3	test4	train2	train3
schedule1	90765	97713	119376	42786	41211	40131
schedule2	27536	23028	12017	7348	12493	3697
schedule3	5784	3023	3108	2217	2589	1119

Table 1: Number of turns under different schedules and different data sets

2 Features

The features are shown in Table 5.

team	entry	slot	schedule	metric	test1	test2	test3	test4	train2	train3
baseline	entry0	all	schedule1	accuracy	0.7748	0.7928	0.9178	0.8233	0.7523	0.9313
team0	entry0	all	schedule1	accuracy	0.8162	0.7971	0.7720	0.8434	0.7837	0.7207
team0	entry1	all	schedule1	accuracy	0.7748	0.7928	0.9178	0.8233	0.7523	0.9313
team1	entry1	all	schedule1	accuracy	0.8853	0.8465	0.9230	0.8479	0.8664	0.9459
team2	entry1	all	schedule1	accuracy	0.8674	0.8844	0.9435	0.8351	0.8461	0.9451
team2	entry2	all	schedule1	accuracy	0.8673	0.8832	0.9432	0.8346	0.8454	0.9348
team3	entry1	all	schedule1	accuracy	0.8585	0.8752	0.9458	0.7279	0.8462	0.9440
team3	entry2	all	schedule1	accuracy	0.8661	0.8687	0.9315	0.8631	0.8469	0.9274
team3	entry3	all	schedule1	accuracy	0.8814	0.8836			0.8612	
team4	entry1	all	schedule1	accuracy	0.7975	0.8445	0.9344	0.6475	0.7947	0.9540
team5	entry1	all	schedule1	accuracy	0.8576	0.8389	0.9304	0.8500	0.8886	0.9539
team5	entry2	all	schedule1	accuracy	0.8577	0.8438	0.9222	0.8681	0.8920	0.9509
team5	entry3	all	schedule1	accuracy	0.7301	0.7677	0.9013		0.8850	0.9429
team5	entry4	all	schedule1	accuracy	0.8530	0.8563	0.9118		0.8887	0.9502
team5	entry5	all	schedule1	accuracy	0.8649	0.8848	0.9045		0.8857	0.9537
team6	entry1	all	schedule1	accuracy	0.9115	0.9240*	0.8424	0.8673	0.9085*	0.8855
team6	entry2	all	schedule1	accuracy	0.8874	0.9047	0.9481	0.8428	0.8808	0.9700*
team6	entry3	all	schedule1	accuracy	0.9170	0.9202	0.9386	0.8678	0.9045	0.9656
team6	entry4	all	schedule1	accuracy	0.9171*	0.9221	0.9408	0.8672	0.9033	0.9660
team6	entry5	all	schedule1	accuracy	0.8888	0.9043	0.9486	0.8457	0.8784	0.9695
team7	entry1	all	schedule1	accuracy	0.8440	0.8548	0.9224	0.7657	0.8129	0.9333
team8	entry1	all	schedule1	accuracy	0.8283	0.8155	0.8067	0.8289	0.8380	0.8141
team8	entry2	all	schedule1	accuracy	0.8200	0.7982	0.8067	0.8090	0.8571	0.8117
team8	entry3	all	schedule1	accuracy	0.5495	0.6606	0.9214	0.7870	0.6118	0.9523
team8	entry4	all	schedule1	accuracy	0.7800	0.8088	0.9131	0.7973	0.7833	0.9383
team8	entry5	all	schedule1	accuracy	0.7684	0.8087	0.9102	0.8128	0.7662	0.9383
team9	entry1	all	schedule1	accuracy	0.8770	0.8725	0.9441	0.8690	0.8842	0.9679
team9	entry2	all	schedule1	accuracy	0.8732	0.8711	0.9437	0.8701	0.8805	0.9662
team9	entry3	all	schedule1	accuracy	0.8821	0.8825	0.9479	0.8466	0.8822	0.9644
team9	entry4	all	schedule1	accuracy	0.8798	0.8844	0.9487*	0.8415	0.8843	0.9627
team9	entry5	all	schedule1	accuracy	0.8286	0.8276	0.8918	0.8802*	0.8427	0.9455

Table 2: Accuracy of all entries on schedule 1

3 TODO

- Get the Upper Bound of N-Best

team	entry	slot	schedule	metric	test1	test2	test3	test4	train2	train3
baseline	entry0	all	schedule2	accuracy	0.6020	0.4905	0.6202	0.5841	0.5485	0.6641
team0	entry0	all	schedule2	accuracy	0.7056	0.5267	0.2590	0.6585	0.6097	0.2440
team0	entry1	all	schedule2	accuracy	0.6020	0.4905	0.6202	0.5841	0.5485	0.6641
team1	entry1	all	schedule2	accuracy	0.7686	0.6011	0.5948	0.6787	0.7347	0.6760
team2	entry1	all	schedule2	accuracy	0.7167	0.6649	0.6424	0.6196	0.6937	0.6960
team2	entry2	all	schedule2	accuracy	0.7165	0.6617	0.6418	0.6183	0.6922	0.6889
team3	entry1	all	schedule2	accuracy	0.7221	0.6564	0.6453	0.5328	0.7044	0.7235
team3	entry2	all	schedule2	accuracy	0.7186	0.6135	0.5629	0.6826	0.6853	0.7062
team3	entry3	all	schedule2	accuracy	0.7643	0.6735				0.6394
team4	entry1	all	schedule2	accuracy	0.5803	0.5428	0.5699	0.2654	0.5975	0.6668
team5	entry1	all	schedule2	accuracy	0.7654	0.6165	0.6639	0.6448	0.8065	0.7812
team5	entry2	all	schedule2	accuracy	0.7650	0.6255	0.6405	0.6956	0.8097	0.7841^{*}
team5	entry3	all	schedule2	accuracy	0.4688	0.3813	0.5689		0.7941	0.7506
team5	entry4	all	schedule2	accuracy	0.7557	0.6312	0.5897		0.8061	0.7544
team5	entry5	all	schedule2	accuracy	0.7757	0.6958	0.5791		0.7972	0.7614
team6	entry1	all	schedule2	accuracy	0.8172	0.7784^{*}	0.4367	0.6967	0.8155^{*}	0.5926
team6	entry2	all	schedule2	accuracy	0.7593	0.7206	0.6740	0.6293	0.7608	0.7833
team6	entry3	all	schedule2	accuracy	0.8220	0.7586	0.6321	0.6918	0.8076	0.7601
team6	entry4	all	schedule2	accuracy	0.8223^{*}	0.7619	0.6418	0.6885	0.8052	0.7601
team6	entry5	all	schedule2	accuracy	0.7630	0.7190	0.6799^{*}	0.6358	0.7566	0.7801
team7	entry1	all	schedule2	accuracy	0.6880	0.5669	0.5399	0.5078	0.6151	0.6494
team8	entry1	all	schedule2	accuracy	0.7226	0.5733	0.3155	0.6260	0.7212	0.4444
team8	entry2	all	schedule2	accuracy	0.7048	0.5386	0.3031	0.5870	0.7559	0.4252
team8	entry3	all	schedule2	accuracy	0.2365	0.1884	0.6162	0.5275	0.3137	0.7698
team8	entry4	all	schedule2	accuracy	0.6283	0.5224	0.6027	0.5378	0.6211	0.7560
team8	entry5	all	schedule2	accuracy	0.6097	0.5193	0.6024	0.5542	0.5902	0.7560
team9	entry1	all	schedule2	accuracy	0.7824	0.6736	0.6568	0.7050	0.7772	0.7771
team9	entry2	all	schedule2	accuracy	0.7776	0.6720	0.6456	0.7030	0.7733	0.7590
team9	entry3	all	schedule2	accuracy	0.7848	0.6943	0.6413	0.6498	0.7760	0.7222
team9	entry4	all	schedule2	accuracy	0.7820	0.6999	0.6381	0.6399	0.7778	0.7108
team9	entry5	all	schedule2	accuracy	0.7265	0.5887	0.4837	0.7368^{*}	0.7091	0.6746

Table 3: Accuracy of all entries on schedule 2

team	entry	slot	schedule	metric	test1	test2	test3	test4	train2	train3
baseline	entry0	all	schedule2	accuracy	0.5982	0.4869	0.7033	0.6396	0.5647	0.7480
team0	entry0	all	schedule3	accuracy	0.6255	0.4790	0.1200	0.6234	0.5674	0.0822
team0	entry1	all	schedule3	accuracy	0.5982	0.4869	0.7033	0.6396	0.5647	0.7480
team1	entry1	all	schedule3	accuracy	0.7818	0.6381	0.7095	0.6261	0.7482	0.8195
team2	entry1	all	schedule3	accuracy	0.7768	0.7641	0.7973	0.6576	0.7594	0.8320
team2	entry2	all	schedule3	accuracy	0.7765	0.7605	0.7947	0.6558	0.7578	0.8186
team3	entry1	all	schedule3	accuracy	0.7476	0.7228	0.7938	0.3112	0.7300	0.8213
team3	entry2	all	schedule3	accuracy	0.7590	0.7152	0.7590	0.7028	0.7319	0.7945
team3	entry3	all	schedule3	accuracy	0.7877	0.7469			0.7509	
team4	entry1	all	schedule3	accuracy	0.6912	0.6705	0.7722	0.2950	0.6987	0.8463
team5	entry1	all	schedule3	accuracy	0.7284	0.6166	0.7597	0.6748	0.7837	0.8570
team5	entry2	all	schedule3	accuracy	0.7294	0.6325	0.7301	0.7100	0.7937	0.8472
team5	entry3	all	schedule3	accuracy	0.5820	0.4803	0.6409		0.7791	0.8436
team5	entry4	all	schedule3	accuracy	0.6990	0.6414	0.6831		0.7791	0.8391
team5	entry5	all	schedule3	accuracy	0.7438	0.7397	0.6577		0.7806	0.8579
team6	entry1	all	schedule3	accuracy	0.8437	0.8544*	0.4196	0.6775	0.8304	0.6318
team6	entry2	all	schedule3	accuracy	0.8276	0.8131	0.8214	0.6852	0.8080	0.8999
team6	entry3	all	schedule3	accuracy	0.8620	0.8445	0.7857	0.7221	9240*	0.8432
team6	entry4	all	schedule3	accuracy	0.8622*	0.8498	0.7944	0.7212	0.8428	0.8803
team6	entry5	all	schedule3	accuracy	0.8276	0.8121	0.8234	0.6901	0.8073	0.9008*
team7	entry1	all	schedule3	accuracy	0.6964	0.6891	0.6239	0.5016	0.6628	0.7042
team8	entry1	all	schedule3	accuracy	0.6888	0.5577	0.2407	0.6216	0.6914	0.3718
team8	entry2	all	schedule3	accuracy	0.6739	0.5147	0.2416	0.5918	0.7563	0.3753
team8	entry3	all	schedule3	accuracy	0.3119	0.2455	0.7156	0.5498	0.3774	0.8570
team8	entry4	all	schedule3	accuracy	0.6129	0.5471	0.6856	0.5760	0.6099	0.8329
team8	entry5	all	schedule3	accuracy	0.5877	0.5468	0.6766	0.6139	0.5759	0.8329
team9	entry1	all	schedule3	accuracy	0.7827	0.7033	0.8034	0.7262	0.7806	0.8928
team9	entry2	all	schedule3	accuracy	0.7780	0.7049	0.8057	0.7258	0.7783	0.8865
team9	entry3	all	schedule3	accuracy	0.7894	0.7321	0.8240	0.6951	0.7771	0.8829
team9	entry4	all	schedule3	accuracy	0.7872	0.7476	0.8279*	0.6816	0.7876	0.8758
team9	entry5	all	schedule3	accuracy	0.6712	0.5703	0.6116	0.7406*	0.6956	0.8213

Table 4: Accuracy of all entries on schedule 3

Team	Features
Henderson et al.	SLU score
	Rank score
	Affirm score
	Negate score
	Go back score
	Implicit score
	User act type
	Machine act type
	Cant help
	Slot confirmed
	Slot requested
	Slot informed
Ren et al.	Pairwise-slots of the same rank
	Pairwise-slots with identical value
	SLU score and rank of slot
	Dialog history (grounding information)
	Count of slots with identical value
	Domain-specific features
	Baseline Tracker
Metallinou et al.	rank of the current SLU result
	the SLU result confidence score(s)
	the difference between the current hypothesis score and the best
	the number of times an SLU result has been observed before
	the number of times an SLU result has been observed before at a specific rank
	the sum and average of confidence scores
	the number of possible past user negations or confirmations
Lee 1	$informs_k(y, x_1^t)$
	$affirm_k(y, x_1^t)$
	$max\_score_k(y, x_1^t)$
	$acc\_score(y, x_1^t)$
	$pbm\_score(y, x_1^t)$
	$prior_k(y, x_1^t)$
	$canthelpk(y, x_1^t)$
	$bias(y, x_1^t)$
	$bias_{none}(y, x_1^t)$

Table 5: Features