THE FIFTH DIALOG STATE TRACKING CHALLENGE

Seokhwan Kim[1], Luis Fernando D'Haro[1], Rafael E. Banchs[1], Jason D. Williams[2], Matthew Henderson[3], Koichiro Yoshino[4]
[1]Institute for Infocomm Research, Singapore. [2]Microsoft Research, USA.
[3]Google, USA. [4]Nara Institute of Science and Technology, Japan.

ABSTRACT

Dialog state tracking – the process of updating the dialog state after each interaction with the user – is a key component of most dialog systems. Following a similar scheme to the fourth dialog state tracking challenge, this edition again focused on human-human dialogs, but introduced the task of cross-lingual adaptation of trackers. The challenge received a total of 32 entries from 9 research groups. In addition, several pilot track evaluations were also proposed receiving a total of 16 entries from 4 groups. In both cases, the results show that most of the groups were able to outperform the provided baselines for each task.

Index Terms— Dialog state tracking, cross-lingual adaptation, challenge report.

1. INTRODUCTION

Dialog state tracking refers to the task of updating the dialog state after each interaction between the system and the user, where the dialog state represents the history of the conversation up to the current timestep. Since most dialog managers choose their next action based wholly or at least in part on the dialog state, dialog state tracking is one of the key problems in building a dialog system. Yet dialog state tracking is difficult because errors in speech recognition and language understanding render the true state of the dialog partially observable, and because natural language can be ambiguous.

The Dialog State Tracking Challenge was created to promote research in this area and to provide a common benchmark for evaluating this task. Four different challenges have been organized previously [1, 2, 3, 4]. Like in the fourth challenge, in this fifth edition we also focused on state tracking for human-human dialogs that has the desirable property that no pre-specified policy is followed, yielding dialogs that are varied. This approach hedges against building a dialog state tracking that over-fits to a particular fixed dialog manager, and takes a step toward building dialog systems directly from a corpus of human-human interactions. For this challenge, dialog states are defined at the sub-dialog segment level using a frame structure that consists of slot-value pairs representing the main subject in each sub-dialog. Therefore, trackers are required to fill out the frames by considering the previous dialog turns in a given sub-dialog.

Different to the previous (fourth) challenge, in this edition we introduced a cross-language dialog state tracking task, aiming at addressing the problem of adaptation to a new language. In this case, the goal is to build a tracker for the target language using existing resources in the source language and the corresponding machine translated sentences in the target language. In addition to the main task, we also proposed different pilot tracks for the core components in developing end-to-end dialog systems following the same cross-language setting. Our goal with this new focus is to contribute to the progress on improving the language portability of state-of-the-art monolingual technologies while reducing the costs of developing dialog systems for resource-poor target languages.

2. CHALLENGE OVERVIEW

2.1. Challenge design

This fifth challenge shares much in common with DSTC4 [4] in the definitions of the target tasks and the characteristics of the datasets. This section gives a summary of the challenge and the newly introduced cross-language aspects of the tasks.

2.1.1. Main Task

The main task aims at tracking the dialog state defined as a frame structure filled with slot-value pairs representing the subject of each sub-dialog in human-human dialogs. Sub-dialogs have been manually segmented from a full dialog session and annotated with the topic category. For each turn in a given sub-dialog, its topic-specific frame should be filled out considering all the dialog history up to the turn. Fig. 1 shows examples of Chinese dialog segments annotated with their dialog state frames.

Different from DSTC4, this challenge addresses a cross-language dialog state tracking problem. Firstly, a training set of labelled dialogs in English and a small development set in Chinese were released to participants at the beginning of the challenge. During the developing phase, the participants built their systems for state tracking in Chinese dialogs using the English training set. Finally, the performance of each tracker was evaluated on the unlabelled test set in Chinese by comparing the system outputs with reference annotations.

Speaker	Utterance	Dialog State
Guide	我介绍你这个甘榜格南。 (I recommend you this Kampong Glam.)	**TOPIC: Attraction** TYPE OF PLACE: *Ethnic enclave* NBHD: *Kampong Glam*
Tourist	对。(Right.)	
Guide	你看,它是个-它是马来村嘛 (You see, it is a- it's a Malay Village)	
Tourist	对，甘榜- (Right, Kampong-)	
Guide	它就卖了很多马来食物。 (It sells a lot of Malay food.)	**TOPIC: Food** CUISINE: *Malay cuisine* NEIGHBORHOOD: *Kampong Glam*
Tourist	比较有特色的食物, (It's quite a unique food,)	
Guide	对，哦。(Right.)	
Guide	马来食物，基本上，它是香。 (Malay food, basically, it smells very nice.)	
Tourist	那我们住宿呢？(Then, where do we stay?)	
Guide	我介绍一间呵，叫V Hotel的。 (Let me recommend to you, the V Hotel.)	**TOPIC: Accommodation** INFO: *Pricerange* NAME: *V Hotel*
Guide	这个酒店，价格这个不贵。 (This hotel, the price is not expensive.)	
Tourist	好的。(Okay.)	
Guide	如果要去，我建议的这个马来文化村， (If you want to go, I suggest this Malay cultural village,)	**TOPIC: Transportation** INFO: *Duration* TYPE: *Walking* FROM: *V Hotel* TO: *Kampong Glam*
Tourist	马来村? (Malay village?)	
Guide	步行大概我看十五分钟吧。 (I think it would take fifteen minutes on foot.)	
Tourist	好。(That's good.)	

Fig. 1: Example human-human dialog in Chinese and dialog state labels for the main task of DSTC5.

2.1.2. Pilot Tasks

As in DSTC4, this challenge also included these pilot tasks:

- Spoken language understanding (SLU): Tagging a given utterance with speech acts and semantic slots.
- Speech act prediction (SAP): Predicting the speech act of the next turn imitating the policy of one speaker (tourist or guide).
- Spoken language generation (SLG): Generating a response utterance for one speaker by using the corresponding speech act and semantic slot information.
- End-to-end system (EES): Developing an end-to-end system by pipe- lining and/or combining different SLU, SAP and SLG systems[1].

Following the same cross-language scenario as in the main task, each system for the pilot tasks was to be trained on English dialogs and then evaluated over Chinese dialogs. Further information can be found in DSTC5 handbook [5].

2.2. Data

The challenge uses the TourSG corpus [5], which consists of dialog in both English and Chinese about tourist information for Singapore collected from Skype calls between tour guides and tourists. All the recorded dialogs, with a total duration of 21 hours per language, were manually transcribed and annotated with speech act and semantic labels at the turn level.

[1]As in DSTC4, no teams opted to participate in the end-to-end system track. We conjecture this is because of the substantial effort required to field an end-to-end system.

Table 1: Overview of DSTC5 data

Set	Task	Language	# dialogs	# utterances
Train	ALL	English	35	31,304
Dev	ALL	Chinese	2	3,130
Test	MAIN	Chinese	10	14,878
Test	SLU	Chinese	8	12,655
Test	SAP	Chinese	8	11,456
Test	SLG	Chinese	8	12,346

The whole English part of the dataset, which was used previously in DSTC4, was released as training set for all the tasks in DSTC5. The Chinese dialog set was divided into five subsets, including the development set (with just two sessions) and four test sets corresponding to the main task and three pilot tasks (Table 1). Each dialog session in both the training and the test sets for the main task was manually segmented and annotated with one of the five major topic categories and its corresponding frame structure.

In addition to the original dialogs, 5-best translations were provided for each utterance. These translations were generated by using an English-to-Chinese machine translation system for the training set and a Chinese-to-English system for the development and test sets. All translated utterances were given with the corresponding word alignment information.

Along with the dialog corpus, a ontology was also provided to describe the tagset definitions as well as the domain knowledge regarding tourism in Singapore. The original DSTC4 ontology, which was created based on the English dialogs, have been expanded with additional contents occurring in the Chinese dialogs. It also included 5-best results of English-to-Chinese machine translation for each entry.

3. EVALUATION

3.1. Main Task

3.1.1. Evaluation metrics

Main task systems were required to generate a tracking output for every turn in a given log file. Trackers were allowed to use all the transcriptions and sub-dialog details provided in the log object from the beginning of the session, up to the current turn. Trackers were prohibited from using any information from the future turns, because in a practical dialog system, this information would not yet be available.

To examine the capabilities of a tracker for both understanding the contents in a given sub-dialog and predicting its dialog states, two different schedules were considered to select the utterances for the target of evaluation:

- Schedule 1: all turns are included
- Schedule 2: only the turns at the end of sub-dialogs are included

The following metrics were used for evaluation:

- Accuracy: Fraction of sub-dialogs in which the tracker's output is equivalent to the gold standard frame structure
- Precision: Fraction of slot-value pairs in the tracker's outputs correctly filled
- Recall: Fraction of slot-value pairs in the gold standard labels correctly filled
- F-measure: The harmonic mean of precision and recall

3.1.2. Baseline tracker

A simple baseline tracker was provided for the main task. It is based on the rule-based system used in DSTC4, which determines the slot values by fuzzy string matching between the entries in the ontology and the transcriptions of the utterances mentioned from the beginning of a given sub-dialog to the current turn. To adapt it for the cross-language execution, the following two different methods were implemented.

- Method 1: The translated utterances from Chinese-to-English are matched to the English entries in the original ontology (team 0 / entry 0).
- Method 2: The Chinese utterances are matched to the translated entries in the ontology from English-to-Chinese (team 0 / entry 1).

For both methods, only the top-1 hypothesis of the 5-best translations was used for each matching. If a part of a given utterance was matched with an entry for a slot in the ontology with over a certain level of similarity, the entry was simply assigned as a value for the particular slot in the tracker's output.

3.1.3. Results

In total 32 entries were submitted from 9 research teams. To preserve anonymity, the teams were identified by numbers from 1 to 9. The baseline system was marked as team 0.

Table 2 shows the averaged results over the whole test set for each submitted entry. More specific scores by topic and slot type and all the submitted entries are available on the DSTC5 repository [2]. As seen from the table, most of the trackers outperformed the baseline in all the combinations of schedules and metrics. Especially, the best entries from *team 2* achieved almost three times and more than twice as high performances as the baseline in accuracy and F-measure,

[2] https://github.com/seokhwankim/dstc5

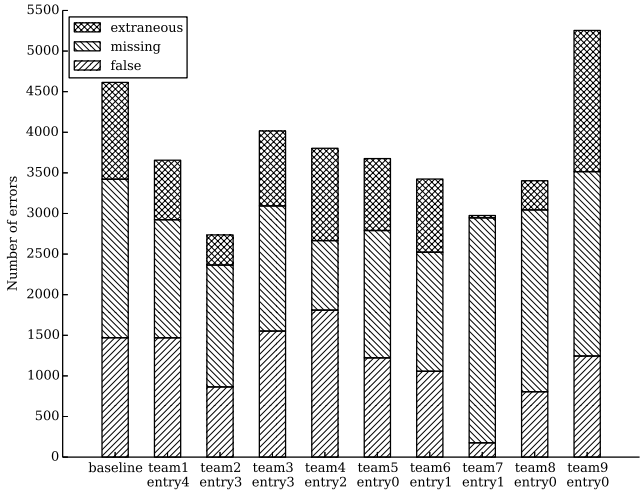

Fig. 3: Number of errors made by type for the best entry from each team in the main task on the test set.

respectively, under both schedules. Fig. 2 reveals that the highly-ranked trackers in the overall comparison tend to produce evenly good results across all topic categories. To investigate the reasons for the performance differences among the trackers, the slot-level errors under Schedule 2 from the best entry of each team were categorized into the three error types following [6]:

- Missing attributes: if the reference contains values for a slot, but the tracker outputs no value for that slot.
- Extraneous attributes: if the reference contains no value for a slot, but the tracker outputs values for that slot.
- False attributes: if the reference contains values for a slot, but the tracker outputs wrong values for that slot.

The error distributions in Fig. 3 indicate that *team2.entry3* achieved the higher performances by much lower numbers of extraneous and false attributes than the others. On the other hand, *team4.entry2* yielded competitive results by reducing the missing slot errors.

3.2. Pilot Tasks: SLU and SAP

3.2.1. Evaluation metrics

Each pilot task includes two subtasks, one each for modelling the tour guide and the tourist. Both SLU and SAP tasks share the same utterance-level annotations for speech acts and semantic tags. In the SLU task, a system is required to produce both semantic tags and speech acts for a given unlabelled utterance spoken by the target role speaker. For the SAP task, an input included the utterance from the other speaker labelled with both speech acts and semantic tags along with the resulting semantic tags for the next turn by the target speaker. Then,

Table 2: Main task results on the test set. Team 0 is the rule-based baseline. Bold denotes the best result in each column.

		Schedule 1				Schedule 2			
Team	Entry	Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
0	0	0.0250	0.1148	0.1102	0.1124	0.0321	0.1425	0.1500	0.1462
0	1	0.0161	0.1743	0.1279	0.1475	0.0222	0.1979	0.1774	0.1871
1	0	0.0397	0.3320	0.2934	0.3115	0.0551	0.3429	0.3712	0.3565
1	1	0.0386	0.3615	0.2610	0.3032	0.0597	0.3785	0.3324	0.3540
1	2	0.0393	0.3673	0.2639	0.3071	0.0551	0.3794	0.3358	0.3563
1	3	0.0387	0.4087	0.2436	0.3052	0.0597	0.4260	0.3087	0.3580
1	4	0.0417	0.3650	0.2795	0.3166	0.0612	0.3811	0.3548	0.3675
2	0	0.0736	0.4664	0.3449	0.3966	**0.0964**	0.5217	0.3849	0.4430
2	1	0.0567	0.3712	0.3818	0.3764	0.0712	0.4340	0.4196	0.4267
2	2	0.0529	0.3629	**0.3892**	0.3756	0.0681	0.4216	**0.4303**	0.4259
2	3	**0.0788**	0.5195	0.3315	**0.4047**	0.0956	**0.5643**	0.3769	**0.4519**
2	4	0.0699	0.4862	0.3432	0.4024	0.0872	0.5427	0.3842	0.4499
3	0	0.0351	0.3216	0.1515	0.2060	0.0505	0.3350	0.2045	0.2539
3	1	0.0303	0.2648	0.2235	0.2424	0.0367	0.2788	0.2873	0.2830
3	2	0.0289	0.3182	0.1538	0.2074	0.0406	0.3377	0.2078	0.2573
3	3	0.0341	0.2926	0.2095	0.2442	0.0451	0.3076	0.2733	0.2895
4	0	0.0583	0.4008	0.2776	0.3280	0.0765	0.4127	0.3284	0.3658
4	1	0.0407	0.3554	0.3267	0.3405	0.0413	0.3569	0.3575	0.3572
4	2	0.0515	0.3682	0.3735	0.3708	0.0635	0.3768	0.4140	0.3945
4	3	0.0552	0.3717	0.3583	0.3649	0.0681	0.3806	0.4026	0.3913
4	4	0.0454	0.3473	0.3677	0.3572	0.0559	0.3510	0.4043	0.3758
5	0	0.0330	0.3377	0.2318	0.2749	0.0520	0.3637	0.3044	0.3314
5	1	0.0187	0.1474	0.2325	0.1804	0.0230	0.1611	0.2526	0.1967
5	2	0.0183	0.1973	0.1236	0.1520	0.0168	0.2003	0.1042	0.1371
5	3	0.0313	0.1506	0.1648	0.1574	0.0413	0.1728	0.2062	0.1880
5	4	0.0093	0.4265	0.0531	0.0945	0.0115	0.4286	0.0551	0.0977
6	0	0.0389	0.4467	0.2092	0.2849	0.0482	0.4509	0.2516	0.3230
6	1	0.0340	0.3897	0.2533	0.3070	0.0383	0.4063	0.3124	0.3532
6	2	0.0491	0.4684	0.2193	0.2988	0.0643	0.4758	0.2623	0.3381
7	0	0.0092	0.4287	0.0431	0.0783	0.0107	0.4000	0.0441	0.0794
7	1	0.0085	**0.5892**	0.0410	0.0767	0.0115	0.5369	0.0438	0.0809
8	0	0.0192	0.3130	0.1048	0.1570	0.0214	0.3021	0.1046	0.1554
8	1	0.0068	0.0924	0.0395	0.0554	0.0069	0.0948	0.0414	0.0577
9	0	0.0231	0.1139	0.1090	0.1114	0.0314	0.1412	0.1487	0.1449

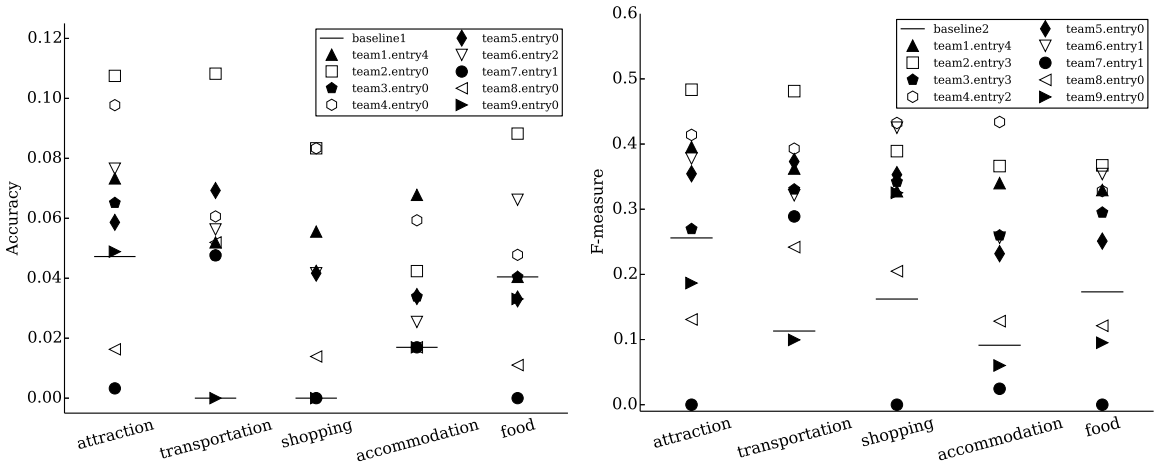

Fig. 2: Comparisons of the main task performances per topic category among the best tracker from each team

Table 3: Results of speech act identification in SLU on the test set.

		Guide			Tourist		
Team	Entry	Precision	Recall	F-measure	Precision	Recall	F-measure
0	0	0.4588	0.2480	0.3219	0.3694	0.1828	0.2446
2	0	0.5450	0.3911	0.4554	0.5001	**0.5501**	0.5239
2	1	0.5305	0.3969	0.4540	0.5331	0.5263	**0.5297**
2	2	**0.5533**	0.3829	0.4526	0.5107	0.5425	0.5261
2	3	0.5127	**0.4251**	**0.4648**	**0.5605**	0.4999	0.5285
3	0	0.4279	0.3583	0.3900	0.4591	0.4241	0.4409
3	1	0.4340	0.3635	0.3956	0.4498	0.4119	0.4300
5	0	0.4085	0.3364	0.3690	0.5026	0.4484	0.4739
5	1	0.3905	0.3216	0.3527	0.4519	0.4031	0.4261
5	2	0.4639	0.3820	0.4190	0.4916	0.4385	0.4635
5	3	0.4540	0.3739	0.4101	0.4871	0.4346	0.4594
5	4	0.4459	0.3672	0.4028	0.4984	0.4446	0.4700
7	0	0.5007	0.2976	0.3733	0.5079	0.4156	0.4571

Table 4: Results of semantic tagging in SLU on the test set.

		Guide			Tourist		
Team	Entry	Precision	Recall	F-measure	Precision	Recall	F-measure
0	0	0.4666	0.3187	**0.3787**	0.5259	0.2659	0.3532
3	0	0.4650	0.3182	0.3779	**0.5331**	0.2620	0.3513
3	1	0.4650	0.3182	0.3779	**0.5331**	0.2620	0.3513
5	0	0.5006	0.2923	0.3691	0.5083	**0.3110**	**0.3859**
5	1	**0.5469**	0.1893	0.2813	0.5121	0.3081	0.3847
5	2	0.3577	0.2476	0.2926	0.3031	0.2237	0.2574
5	3	0.3486	0.2541	0.2939	0.2932	0.2149	0.2480
5	4	0.3395	0.2111	0.2603	0.2947	0.2072	0.2433
7	0	0.4400	**0.3207**	0.3710	0.4408	0.2926	0.3517

the system is expected to generate the speech acts for the next utterance. The following evaluation metrics were used:

- Speech acts for SLU and SAP
  - Precision: Fraction of speech act labels that are correctly predicted.
  - Recall: Fraction of speech act labels in the gold standard that are correctly predicted.
  - F-measure: The harmonic mean between precision and recall
- Semantic tags for SLU
  - Precision: Fraction of correctly predicted semantic tags in the generated tag sequences encoded using BIO scheme.
  - Recall: Fraction of correctly predicted semantic tags in the gold standard tag sequences encoded using BIO scheme.
  - F-measure: The harmonic mean between precision and recall

3.2.2. Baseline system

A simple baseline system was also provided for the cross-language SLU pilot task. It used a pair of support vector machines (SVMs) and conditional random fields (CRFs) models trained with basic bag-of-words features for multi-label speech act prediction and semantic tagging, respectively. The models were trained with the English training dataset. Then, the trained models were used to analyze the English translations of each Chinese utterance in the test set. Finally, the predicted labels were projected into the original utterances in Chinese by means of the word alignment information obtained from the machine translation system.

The baseline for the SAP task used SVM models for multi-label speech act prediction using the following features: the semantic tags in the current and the previous utterances, the speech act tags of the recent utterance spoken by the other speaker, and the distance from the other speaker's turn to the current utterance. Since all these features are language-independent, the model trained on the English training set was directly applied on the Chinese test set.

3.2.3. Results

For the SLU task, 4 teams participated with a total of 12 entries submitted [3]. For the SAP task, no submissions were received. Table 3 and Table 4 present the summary of the SLU results for speech act identification and semantic tagging, respectively. All the submitted entries achieved much better performances than the baseline in speech act identification for both speakers. Especially, all the entries from *team2* outperformed not only the baseline, but also all the other teams.

However, most participants failed to show significant improvements in semantic tagging performances. Only two entries *team5.entry0* and *team5.entry1* produced better results than the baseline and only for the case of the tourist turns.

3.3. Pilot Task: SLG

3.3.1. Evaluation metrics

In this task, the system was intended to receive as input the semantic tags and speech acts from the target speaker only, and it was expected to produce the final surface form of the utterance. Here, the goal was to maximize the syntactic and semantic similarity between the gold standard and the generated sentence. The following similarity metrics were used:

- BLEU: Geometric average of n-gram precision (for n = 1, 2, 3, 4) of the system generated utterance with respect to the reference utterance [7].
- AM-FM: Linear interpolation of (1) the cosine similarity between the system generated utterance and the reference utterance and (2) the normalized n-gram probability of the system generated utterance [8]. In brief, AM measures the semantic similarity between the reference and the generated sentence, while FM measures similarity of the n-gram language model probabilities.

3.3.2. Baseline system

The SLG baseline was based on an example-based language generation approach, which used k-nearest neighbors algorithm on the vector space with the speech act and semantic tags features. For each input, the system searched for the most similar item in the English training set, and then returned the top-1 hypothesis of its Chinese translations as result.

3.3.3. Results

For the SLG task, only one team participated with a total of 4 entries. The results are depicted in Table 5. All the submitted entries achieved better performance than the baseline, especially for generating the tourist sentences. Table 6 provides an example of a sentence generated by the baseline and the corresponding entry from the participant.

[3]Team 2 participated only in speech act identification.

Table 5: SLG results on the test set.

		Guide		Tourist	
Team	Entry	AM-FM	BLEU	AM-FM	BLEU
0	0	0.1981	**0.3854**	0.2602	**0.5921**
5	0	0.2818	0.3264	0.3221	0.4850
5	1	**0.3180**	0.3371	0.3635	0.5249
5	2	0.2737	0.2852	0.3100	0.4741
5	3	0.2405	0.2758	**0.4258**	0.5302

Table 6: Example of generated sentences by the baseline and the best entry from the participant

Input Info
Speech act: FOLLOW(INFO, WHERE)
Semantic tags: ATTRACTION(LOCATION: 圣淘沙 (Sentosa))
Reference
现在你看到的这个就是圣淘沙，啊。
(what you see now, this, it is Sentosa.)
Baseline
就是那个位置在圣淘沙。 (it is that position on Sentosa)
BLEU: 0.280 AM: 0.626 FM: 0.783 AM-FM: 0.705
Best Entry
那是圣淘沙。 (that is Sentosa)
BLEU: 0.103 AM: 0.461 FM: 0.964 AM-FM: 0.713

4. CONCLUSIONS

We have presented the official evaluation results of the Fifth Dialog State Tracking Challenge (DSTC5). This edition has focused on dialog state tracking in human-human dialogs, but introducing the problem of adaptation to a new language. A total of 9 teams participated in the main task with an overall number of 32 entries submitted. In addition, four pilot tasks were proposed that received a total of 12 entries from 4 groups for the Spoken Language Understanding (SLU) and 4 entries from one group for the Spoken Language Generation (SLG) tasks, respectively.

Following the same success of DSTC4, we have confirmed the feasibility of performing dialog state tracking at the sub-dialogue level in the context of human-human dialogs, while simultaneously addressing the problem of handling a cross-lingual setting. Although the reported results were in most cases better than the provided baselines, it is clear that more work is needed to produce high-accuracy trackers in this setting. Additionally,there is the need for increasing the volume of training data to test more complex approaches, such as methods based on deep learning.

Overall, we feel that DSTC5 has been a meaningful step toward the long-term goal of creating modular, data-driven end-to-end dialog systems systems. Based on the sustained level of participation over the DSTC series, and on continuing technical advances, we have begun planning for a sixth edition of the challenge.

5. REFERENCES

[1] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black, "The dialog state tracking challenge," in *Proceedings of the SIGDIAL 2013 Conference*, 2013, pp. 404–413.

[2] Matthew Henderson, Blaise Thomson, and Jason Williams, "The second dialog state tracking challenge," in *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2014, p. 263.

[3] Matthew Henderson, Blaise Thomson, and Jason D Williams, "The third dialog state tracking challenge," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 324–329.

[4] Seokhwan Kim, Luis Fernando D'Haro, Rafael E. Banchs, Jason Williams, and Matthew Henderson, "The Fourth Dialog State Tracking Challenge," in *Proceedings of the 7th International Workshop on Spoken Dialogue Systems (IWSDS)*, 2016.

[5] Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason Williams, Matthew Henderson, and Koichiro Yoshino, "Dialog state tracking challenge 5 handbook," 2016, `https://github.com/seokhwankim/dstc5/raw/master/docs/handbook_DSTC5.pdf`.

[6] Ronnie W Smith, "Comparative error analysis of dialog state tracking," in *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2014, p. 300.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[8] Rafael E Banchs, Luis F D'Haro, and Haizhou Li, "Adequacy–fluency metrics: Evaluating mt in the continuous space model framework," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 3, pp. 472–482, 2015.