# The Fourth Dialog State Tracking Challenge

Seokhwan Kim, Luis Fernando D'Haro, Rafael E. Banchs, Jason D. Williams, and Matthew Henderson

**Abstract** Dialog state tracking is one of the key sub-tasks of dialog management, which defines the representation of dialog states and updates them at each moment on a given on-going conversation. To provide a common test bed for this task, three dialog state tracking challenges have been completed. In this fourth challenge, we focused on dialog state tracking on human-human dialogs. The challenge received a total of 24 entries from 7 research groups. Most of the submitted entries outperformed the baseline tracker based on string matching with ontology contents. Moreover, further significant improvements in tracking performances were achieved by combining the results from multiple trackers. In addition to the main task, we also conducted pilot track evaluations for other core components in developing modular dialog systems using the same dataset.

## 1 Introduction

Dialog systems interact with users using natural language to help them achieve a goal, and are increasingly becoming a part of daily life, with examples including Apple's Siri, Google Now, Xbox and Cortana from Microsoft, Facebook M, among others. As the dialog progresses, the dialog system maintains a representation of the state of the dialog in a process called dialog state tracking (DST). For example, in a travel planning system, the dialog state might indicate the search parameters for the type of hotel the user has said they're searching for, such as their desired star rating,

Seokhwan, Kim, Luis Fernando D'Haro, Rafael E. Banchs
Institute for Infocomm Research, Singapore. e-mail: {kims,luisdhe,rembanchs}@i2r.a-star.edu.sg

Jason D. Williams
Microsoft Research, Redmond, WA, USA. e-mail: jason.williams@microsoft.com

Matthew Henderson
Google e-mail: matt@matthen.com

location, and price range. Dialog state tracking is difficult because automatic speech recognition (ASR) and spoken language understanding (SLU) errors are common, and can cause the system to misunderstand the user. Moreover, it can be difficult to determine when to retain information and at the same time, state tracking is crucial because the system relies on the estimated dialog state to choose actions, for example, which hotels to suggest.

To provide a common test bed for this task, three Dialog State Tracking Challenges (DSTCs) have been organized [9, 2, 3]. Different from the previous challenges which had focused on human-machine dialogs, in this fourth edition, we have focused on dialog state tracking in human-human dialogs. The goal of the main task in this challenge was to track dialog states for sub-dialog segments. For each turn in a given sub-dialog, the tracker was required to fill out a frame of slot-value pairs considering all dialog history prior to the turn. We expect these shared efforts on human dialog state tracking will contribute to progress in developing much more human-like dialog systems.

In addition to the main task, this fourth edition of the challenge also proposed a series of pilot tasks for evaluating each of the core components needed for developing end-to-end dialog systems. More specifically, four pilot tasks were offered: Spoken Language Understanding (SLU), Speech Act Prediction (SAP), Spoken Language Generation (SLG), and End-to-end system (EES). This effort constitutes a first step towards the construction of distributed modular systems and the development of a computational framework for collaborative end-to-end system evaluation. In the evaluation, one team participated in the SLU pilot task, but all the data and tasks remain available for research use.

The rest of the paper is organized as follows. Section 2 provides a general overview of the challenge tasks and the used dataset. Sections 3, 4 and 5 describes evaluation results of the main task, while section 6 describes evaluation results of the SLU pilot tasks. Finally, section 7 presents our main conclusions and recommendations.

## 2 Challenge Overview

### *2.1 Problem statement*

#### 2.1.1 Main Task

The goal of the main task is to evaluate state tracking for human-human dialogs between tourists and tour guides. Since each subject in these conversations tends to be expressed not just in a single turn, but through a series of multiple turns, a dialog state is defined for each sub-dialog segment level as a frame structure filled with slot-value pairs representing the main subject of a given segment. Fig. 1 shows examples of segment-level dialog state frame structures.

Each frame could have two different kinds of slots: regular slots and *INFO* slot. While regular slots should be filled with particular values explicitly discussed in the segment, *INFO* slots indicate the subjects that are discussed but not directly related to any particular values of other slots. The possible slot types and the list of their candidate values vary by topic category, which are described in an ontology.

In this challenge, a dialog session segmented into a series of sub-dialogs labeled with topic categories is given as an input to a tracker. For each turn in a given sub-dialog, the frame should be filled out considering all dialog history up to the current turn. The performance of a tracker is evaluated by comparing its outputs with reference annotations.

| Speaker | Utterance | Dialog State |
|---|---|---|
| Tourist | Can you give me some uh- tell me some cheap rate hotels, because I'm planning just to leave my bags there and go somewhere take some pictures. | |
| Guide | Okay. I'm going to recommend firstly you want to have a backpack type of hotel, right? | |
| Tourist | Yes. I'm just gonna bring my backpack and my buddy with me. So I'm kinda looking for a hotel that is not that expensive. Just gonna leave our things there and, you know, stay out the whole day. | Type=*Hostel*<br>Pricerange=*Cheap* |
| Guide | Okay. Let me get you hm hm. So you don't mind if it's a bit uh not so roomy like hotel because you just back to sleep. | |
| Tourist | Yes. Yes. As we just gonna put our things there and then go out to take some pictures. | |
| Guide | Okay, um- | |
| Tourist | Hm. | |
| Guide | Let's try this one, okay? | |
| Tourist | Okay. | |
| Guide | It's InnCrowd Backpackers Hostel in Singapore. If you take a dorm bed per person only twenty dollars. If you take a room, it's two single beds at fifty nine dollars. | Name=*InnCrowd Backpackers Hostel*<br>Info=*Pricerange* |
| Tourist | Um. Wow, that's good. | |
| Guide | Yah, the prices are based on per person per bed or dorm. But this one is room. So it should be fifty nine for the two room. So you're actually paying about ten dollars more per person only. | |
| Tourist | Oh okay. That's- the price is reasonable actually. It's good. | |

**Fig. 1** Example human-human dialog and dialog state labels for the main task of DSTC4

### 2.1.2 Pilot Tasks

In addition to the main task, the challenge included a series of optional pilot tracks for the core components in developing end-to-end dialog systems using the same dataset and considering either the information from the tourist or the tour guide. The four proposed tasks were:

- Spoken language understanding (SLU): The objective is to tag a given utterance with speech acts (slot values) and semantic slots.
- Speech act prediction (SAP): The objective is to predict the speech act of the next turn imitating the policy of one speaker. Here, the input to the systems will be the utterances and annotations (semantic tags and speech acts) from a given user (i.e. tourist or guide) along with the resulting semantic tags for the next opposite user (i.e. guide or tourist) utterances, and the system must produce the speech acts for the given user utterances.
- Spoken language generation (SLG): The objective is to generate a response utterance for one of the participants by using the corresponding speech act and semantic slot information.
- End-to-end system (EES) The objective is to develop an end-to-end system by pipelining and/or combining different SLU, SAP and SLG systems. Here, the input to the systems will be the one user utterances and the system must produce the other user utterances.

Different from the main task, in which dialog states are defined at the sub-dialog level and each of the sub-dialogs has a frame structure with slot value pairs to represent the subject discussed within it; in the pilot tasks, annotations are provided at the utterance level and, accordingly, systems must deal with slot value pairs at the utterance level.

## *2.2 Challenge design*

Similar to the previous challenges, both the main and pilot problems are studied as corpus-based tasks with static dialogs. In the development phase, a set of labelled dialogs are released to participants so that they train and optimize their models. And then, the developed components produce the outputs on the unlabelled test set in the evaluation phase. Since every participant uses the same shared datasets for both development and evaluation, the results on the test set can be directly compared to each other.

## *2.3 Data*

The data used in the challenge is TourSG corpus which consists of 35 dialog sessions on touristic information for Singapore collected from Skype calls between three tour guides and 35 tourists. These 35 dialogs sum up to 31,034 utterances and 273,580 words. All the recorded dialogs with the total length of 21 hours were manually transcribed and annotated with speech act and semantic labels for each turn level.

For the main task, each full dialog session was divided into sub-dialog segments considering their topical coherence and then they were categorized by topics. Each sub-dialog assigned to one of the five major topic categories has an additional frame

structure with slot value pairs to represent some more details about the subject discussed within the sub-dialog.

For the challenge, TourSG corpus were divided into four parts (Table 1). Training and development sets consist of manual transcriptions and annotations at both utterance and sub-dialog levels for training and optimizing the trackers, respectively. For the test sets, only manual transcriptions without annotations are provided during the evaluation period.

**Table 1** Overview of DSTC4 data. SG1, SG2, and SG3 are the three tour guides that participated in the data collection and acco (accommodation), attr (attraction), food, shop (shopping), trsp (transportation), and other are topic categories of dialog segments.

| | # dialogs | | | | # segments | | | | | | | # utterances |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Set | SG1 | SG2 | SG3 | Total | acco | attr | food | shop | trsp | other | total | |
| Training | 7 | 7 | 0 | 14 | 187 | 762 | 275 | 149 | 374 | 357 | 2,104 | 12,759 |
| Development | 3 | 3 | 0 | 6 | 94 | 282 | 102 | 67 | 87 | 68 | 700 | 4,812 |
| Test (main) | 3 | 3 | 3 | 9 | 174 | 616 | 134 | 49 | 174 | 186 | 1,333 | 7,848 |
| Test (pilot) | 2 | 2 | 2 | 6 | 126 | 352 | 124 | 49 | 119 | 107 | 877 | 5,615 |

Along with the dialog corpus, an ontology was also created to provide the tagset definitions as well as the domain knowledge regarding tourism in Singapore. While subjects of human-machine conversations are inevitably restricted within the knowledge-base contents used in developing the system, human-human dialogs are much more flexible and broad in terms of the coverage of subjects. To make the resource as general as possible, the entries in the ontology were collected not only from the corpus itself, but also from external knowledge sources. First, the structured information were automatically extracted from the Wikipedia articles related to Singapore and the official website of Singapore Tourism Board. Then, the collected instances were validated by matching with the annotations in the corpus. Finally, all the missing parts in the ontology were completed manually to cover all the subjects discussed in the dialogs.

More detailed information about the data can be found from [4].

## 3 Main Task: Evaluation

### *3.1 Evaluation metrics*

A system for the main task should generate the tracking output for every turn in a given log file. While all the transcriptions and segment details provided in the log object from the beginning of the session to the current turn can be used, any information from the future turns are not allowed to be considered to analyze the state at a given turn.

Although the fundamental goal of this tracking is to analyze the state for each sub-dialog level, the execution should be done at each utterance level regardless of the speaker from the beginning to the end of a given session in sequence. It aims at evaluating the capabilities of trackers not only for understanding the contents mentioned in a given segment, but also for predicting its dialog states even at an earlier turn of the segment.

To examine these both aspects of a given tracker, two different schedules are 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 segments are included

If some information is correctly predicted or recognized at an earlier turn in a given segment and well kept until the end of the segment, it will have higher accumulated scores than the other cases where the same information is filled at a later turn under schedule 1. On the other hand, the results under schedule 2 indicate the correctness of the outputs after providing all the turns of the target segment.

In this task, the following two sets of metrics are used for evaluation:

- Accuracy: Fraction of segments in which the tracker's output is equivalent to the gold standard frame structure
- Precision/Recall/F-measure
  - 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

While the first metric is to check the equivalencies between the outputs and the references at the whole frame level, the others can show the partial correctness at each slot-value level.

### *3.2 Baseline tracker*

A simple baseline tracker is provided to participants. The baseline tracker determines the slot values by fuzzy string matching [1] between the entries in the ontology and the transcriptions of the utterances mentioned from the beginning of a given segment to the current turn. If a part of given utterances is matched with an entry for a slot in the ontology with over a certain level of similarity, the entry is simply assigned as a value for the particular slot in the tracker's output. Since this baseline does not consider any semantic or discourse aspects from given dialogs, its performance is very limited and there is much room for improvement.

[1] `https://github.com/seatgeek/fuzzywuzzy`

## 4 Main Task: Results

Logistically, the training and development datasets, the ontology, and the scoring scripts were released to the participants on 15 April 2015. Then, the unlabelled test set for the main task was released on 17 August 2015. In this challenge, a web-based competition platform [2] was newly introduced for receiving submissions and evaluating them automatically. Once an entry was uploaded to the site, the evaluation results were immediately provided to the participant and compared to the others by posting them to the leaderboard page.

Teams were given two weeks to run their trackers on the test set and enter the outputs to the submission system. Following the tradition of the previous challenges, the number of entries submitted by each team was limited up to five. And also, all the results posted on the leaderboard were anonymized. After the evaluation phase, the test labels were released to the participants.

In total, 24 entries were submitted from 7 research teams participating in the main task. Teams were identified by anonymous team numbers team 1-7, and 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 DSTC4 website [3] and the full details on the trackers themselves are published in individual papers at IWSDS 2016. Most submitted trackers outperformed the baseline in all the combinations of schedules and metrics. Especially, the best entries from team 3 achieved more than three times and almost twice as high performances as the baseline in accuracy and F-measure, 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. The entry *team3.entry3* is ranked the best for all the topics except just one, and *team4.entry3* also yields competitive results in all the cases.

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 [8]:

- Missing attributes: when the reference contains values for a slot, but the tracker does not output any value for the slot
- Extraneous attributes: when the reference does not contain any value for a slot, but the tracker outputs values for that slot
- False attributes: when the reference contains values for a slot, and the tracker outputs an incorrect value for that slot

The error distributions in Fig. 3 indicate that the missing slot errors act as a decisive factor in performance variations across teams.

The influences of these false negatives to the tracking performances are demonstrated also in the analysis of correct outputs. Fig. 4 compares the distributions of the

[2] `https://www.codalab.org/competitions/4971`

[3] `http://www.colips.org/workshop/dstc4/results.html`

**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 | 0.0374 | 0.3589 | 0.1925 | 0.2506 | 0.0488 | 0.3750 | 0.2519 | 0.3014 |
| 1 | 0 | 0.0456 | 0.3876 | 0.3344 | 0.3591 | 0.0584 | 0.4384 | 0.3377 | 0.3815 |
| | 1 | 0.0374 | 0.4214 | 0.2762 | 0.3336 | 0.0584 | 0.4384 | 0.3377 | 0.3815 |
| | 2 | 0.0372 | 0.4173 | 0.2767 | 0.3328 | 0.0575 | 0.4362 | 0.3377 | 0.3807 |
| | 3 | 0.0371 | 0.4179 | 0.2804 | 0.3356 | 0.0584 | 0.4384 | 0.3426 | 0.3846 |
| 2 | 0 | 0.0487 | 0.4079 | 0.2626 | 0.3195 | 0.0671 | 0.4280 | 0.3257 | 0.3699 |
| | 1 | 0.0467 | 0.4481 | 0.2655 | 0.3335 | 0.0671 | 0.4674 | 0.3275 | 0.3851 |
| | 2 | 0.0478 | 0.4523 | 0.2623 | 0.3320 | 0.0706 | 0.4679 | 0.3226 | 0.3819 |
| | 3 | 0.0489 | 0.4440 | 0.2703 | 0.3361 | 0.0697 | 0.4634 | 0.3335 | 0.3878 |
| 3 | 0 | **0.1212** | 0.5393 | 0.4980 | 0.5178 | **0.1500** | 0.5569 | 0.5808 | 0.5686 |
| | 1 | 0.1210 | 0.5449 | 0.4964 | 0.5196 | **0.1500** | 0.5619 | 0.5787 | 0.5702 |
| | 2 | 0.1092 | 0.5304 | **0.5031** | 0.5164 | 0.1316 | 0.5437 | **0.5875** | 0.5648 |
| | 3 | 0.1183 | **0.5780** | 0.4904 | **0.5306** | 0.1473 | 0.5898 | 0.5678 | **0.5786** |
| 4 | 0 | 0.0887 | 0.5280 | 0.3595 | 0.4278 | 0.1072 | 0.5354 | 0.4273 | 0.4753 |
| | 1 | 0.0910 | 0.5314 | 0.3122 | 0.3933 | 0.1055 | 0.5325 | 0.3623 | 0.4312 |
| | 2 | 0.1009 | 0.5583 | 0.3698 | 0.4449 | 0.1264 | 0.5666 | 0.4455 | 0.4988 |
| | 3 | 0.1002 | 0.5545 | 0.3760 | 0.4481 | 0.1212 | 0.5642 | 0.4540 | 0.5031 |
| 5 | 0 | 0.0309 | 0.2980 | 0.2559 | 0.2754 | 0.0392 | 0.3344 | 0.2547 | 0.2892 |
| | 1 | 0.0268 | 0.3405 | 0.2014 | 0.2531 | 0.0401 | 0.3584 | 0.2632 | 0.3035 |
| | 2 | 0.0309 | 0.3039 | 0.2659 | 0.2836 | 0.0392 | 0.3398 | 0.2639 | 0.2971 |
| 6 | 0 | 0.0421 | 0.4175 | 0.2142 | 0.2831 | 0.0541 | 0.4380 | 0.2656 | 0.3307 |
| | 1 | 0.0478 | 0.5516 | 0.2180 | 0.3125 | 0.0654 | 0.5857 | 0.2702 | 0.3698 |
| | 2 | 0.0486 | 0.5623 | 0.2314 | 0.3279 | 0.0645 | **0.5941** | 0.2850 | 0.3852 |
| 7 | 0 | 0.0286 | 0.2768 | 0.1826 | 0.2200 | 0.0323 | 0.3054 | 0.2410 | 0.2694 |
| | 1 | 0.0044 | 0.0085 | 0.0629 | 0.0150 | 0.0061 | 0.0109 | 0.0840 | 0.0194 |

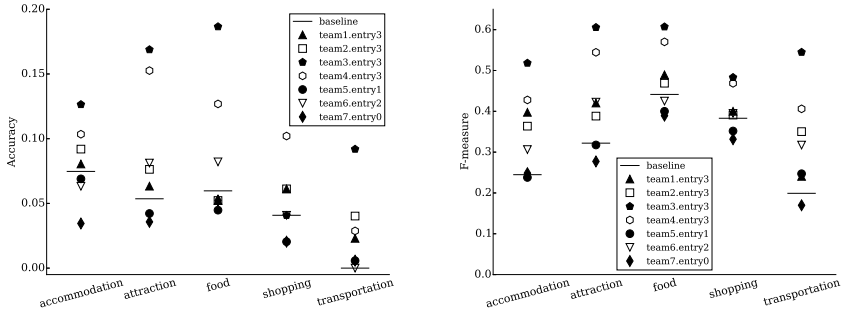


**Fig. 2** Accuracy (left panel) and F-measure (right panel) on the test set per topic for the best tracker from each team in the main dialog state tracking task.

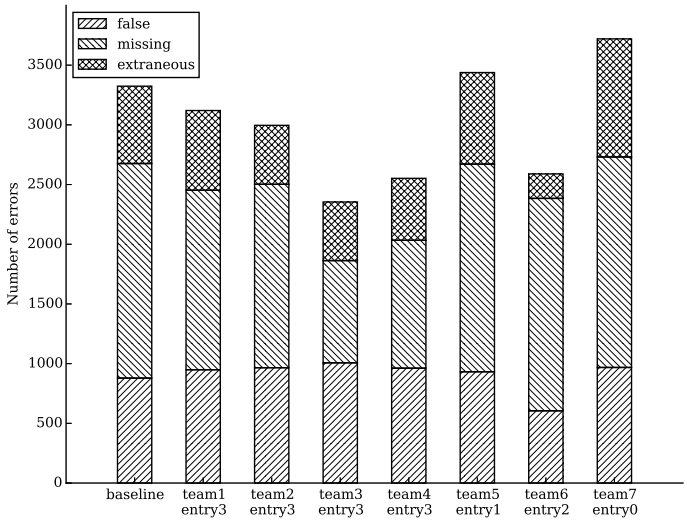


**Fig. 3** Number of errors made by type for the best entry from each team in the main dialog state tracking task on the test set.

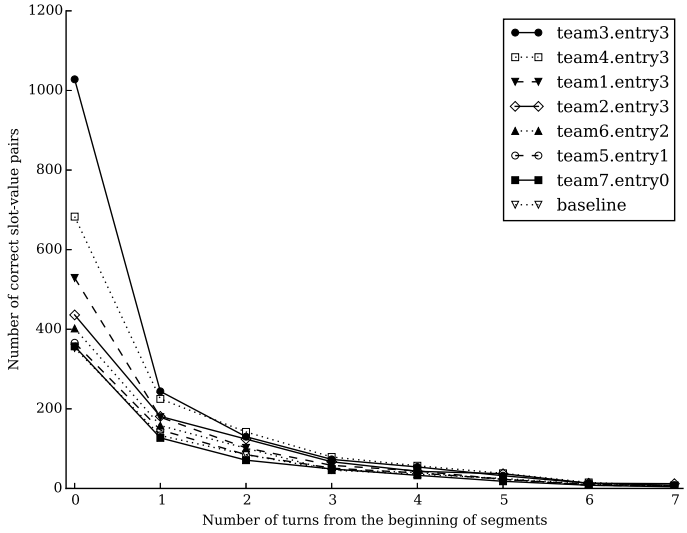


**Fig. 4** Distributions of correct slot-value pairs in the best output from each team by turn offsets where each value is filled for the first time since the beginning of the sub-dialog.

number of correct outputs by the turn offset where each value is filled from the beginning of the sub-dialog. Most of the differences in number of true positives among teams exist at earlier turns of dialog segments, which means that the highly-ranked trackers managed to rescue many slot-value pairs that were missed by others.

## 5 Main Task: Ensemble learning

A merit of corpus-based tasks is that ensemble learning could be studied simply by synthesizing the multiple outputs on the same dataset to improve the performances compared to any single individual system. In the previous dialog state tracking challenges, ensemble learning techniques including score averaging [6] and stacking [2] contributed to improve the tracking performances.

Also for the main task of this challenge, we examined the effectiveness of ensemble learning based on the submitted entries. Since no score information was available in tracking outputs for DSTC4, we adopted the following three simple strategies for combining the outputs:

- Union: fill a slot with a value if the slot-value pair occurs in at least one of the tracking outputs to be combined
- Intersection: fill a slot with a value if the slot-value pair occurs in all the tracking outputs to be combined
- Majority: fill a slot with a value if the slot-value pair occurs in more than half the tracking outputs to be combined

Table 3 compares the performances of combined outputs with the single best entry. The tracking outputs to be combined were selected based on single entry performances in F-measure under Schedule 2 without distinction of team. For example, entry 3, 1, and 0 from team 3 were considered as top 3 entries. The results show that most of the combinations failed to achieve performance improvement from the single best output. Only statistically significant improvement across all metrics was observed when top 3 entries were combined by intersection. This suggests that system combination without considering any correlations among the trackers does not guarantee better results.

To see how much the performances could be improved in case the optimal combination is somehow given considering their correlations, we run the evaluation on every possible combination of 25 entries including the baseline. Table 4 shows the performances of the best combination in each metric. These results are significantly better than the single best entry in most metrics. All the statistical significances in these analyses were computed using approximate randomization [10].

**Table 3** Accuracy and F-measure for various combinations of trackers in the main task on the test set. Bold denotes the best result in each column. +/- indicates statistically significantly better/worse than the single best entry ($p < 0.01$), computed with approximate randomization.

| | Schedule 1 | | Schedule 2 | |
|---|---|---|---|---|
| Tracker | Accuracy | F-measure | Accuracy | F-measure |
| Single best entry | 0.1212 | 0.5306 | 0.1500 | 0.5786 |
| Top 3 entries: union | $0.1111^{-}$ | $0.5147^{-}$ | $0.1325^{-}$ | $0.5619^{-}$ |
| Top 3 entries: intersection | $0.1241^{+}$ | $\mathbf{0.5344}^{+}$ | $\mathbf{0.1561}^{+}$ | $\mathbf{0.5861}^{+}$ |
| Top 3 entries: majority voting | $0.1172^{-}$ | $0.5194^{-}$ | $0.1421^{-}$ | 0.5703 |
| Top 5 entries: union | $0.0980^{-}$ | $0.5133^{-}$ | $0.1107^{-}$ | $0.5543^{-}$ |
| Top 5 entries: intersection | 0.1157 | $0.4370^{-}$ | 0.1369 | $0.5008^{-}$ |
| Top 5 entries: majority voting | $0.1183^{-}$ | $0.5210^{-}$ | 0.1439 | 0.5711 |
| Top 10 entries: union | $0.0623^{-}$ | $0.4719^{-}$ | $0.0680^{-}$ | $0.5014^{-}$ |
| Top 10 entries: intersection | $0.0300^{-}$ | $0.1816^{-}$ | $0.0453^{-}$ | $0.2275^{-}$ |
| Top 10 entries: majority voting | $\mathbf{0.1268}^{+}$ | $0.4741^{-}$ | 0.1456 | $0.5380^{-}$ |
| All entries: union | $0.0077^{-}$ | $0.1320^{-}$ | $0.0078^{-}$ | $0.1366^{-}$ |
| All entries: intersection | $0.0132^{-}$ | $0.0229^{-}$ | $0.0192^{-}$ | $0.0331^{-}$ |
| All entries: majority voting | $0.0646^{-}$ | $0.3535^{-}$ | $0.0898^{-}$ | $0.4135^{-}$ |

**Table 4** The best possible (oracle) combination of trackers in the main task on the test set. All the listed performances were achieved by the majority voting strategy. Bold denotes the best result in each metric. +/- indicates statistically significantly better/worse than the single best entry in Table 3 ($p < 0.01$), computed with approximate randomization.

| | Schedule 1 | | Schedule 2 | |
|---|---|---|---|---|
| Combination | Accuracy | F-measure | Accuracy | F-measure |
| T3E0+T3E2+T3E3+T4E1+T4E3+T6E0+T6E2 | $\mathbf{0.1310}^{+}$ | $0.4870^{-}$ | 0.1517 | $0.5534^{-}$ |
| T3E1+T3E3+T4E2 | $0.1241^{+}$ | $\mathbf{0.5359}^{+}$ | $0.1569^{+}$ | $0.5885^{+}$ |
| T2E3+T3E0+T3E2+T3E3 | $0.1230^{+}$ | $0.5351^{+}$ | $\mathbf{0.1587}^{+}$ | $0.5878^{+}$ |
| T2E3+T3E0+T3E1+T3E2+T3E3+T4E2 | $0.1242^{+}$ | $0.5354^{+}$ | $0.1587^{+}$ | $\mathbf{0.5893}^{+}$ |

# 6 Pilot Tasks

## *6.1 Evaluation metrics*

Two different families of metrics were used for evaluating the pilot tasks: classification accuracy metrics used for SLU and SAP tasks, and semantic similarity metrics used for SLG and EES tasks. For all subtasks in the pilot tasks, evaluation schedule 1 was used (i.e. system outputs are evaluated at all turns). In more detail, the following evaluation metrics were used:

- SLU and SAP tasks:
  - Precision: Fraction of semantic tags and/or speech acts that are correct.
  - Recall: Fraction of semantic tags and/or speech acts in the gold standard that are generated.

  - F-measure: The harmonic mean of precision and recall.
- SLG and EES tasks:
  - 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: Weighted mean 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 [1].

## *6.2 Web-based evaluation*

Regarding operational aspects of pilot task evaluation, participants were required to implement a web-service (WS) to run their systems. During the evaluation, a master evaluation script was used to call the corresponding web-services at specified time slots during the evaluation dates. In order to facilitate these implementations, a server and client python scripts were provided with default configuration to check that the systems were working and reachable from outside local network.

During the evaluation, the participant's server received a JSON object containing the input parameters required for the given task and role and the server used the input parameters to generate a corresponding answer that was send back to the organizer's client using a JSON message. Then, based on the retrieved result, the client calculated the actual values for the proposed metrics. For debugging purposes, both the server and client included a logging module to keep record of all the requests and answers interchanged between both modules. For additional information about the pilot task, messages, and provided scripts please refer to [5].

**Table 5** Results from 5 entries submitted by one team to the NLU task, on the test set.

| | | Speech Act | | | Semantic Tag | | |
|---|---|---|---|---|---|---|---|
| Speaker | Entry | Precision | Recall | F-measure | Precision | Recall | F-measure |
| Guide | 1 | 0.629 | 0.519 | 0.569 | 0.565 | 0.489 | 0.524 |
| | 2 | 0.633 | 0.523 | 0.573 | 0.565 | 0.489 | 0.524 |
| | 3 | **0.745** | **0.615** | **0.674** | 0.565 | 0.489 | 0.524 |
| | 4 | 0.631 | 0.521 | 0.571 | 0.565 | 0.489 | 0.524 |
| | 5 | 0.676 | 0.558 | 0.612 | 0.565 | 0.489 | 0.524 |
| Tourist | 1 | 0.358 | 0.298 | 0.325 | 0.574 | 0.476 | 0.521 |
| | 2 | 0.293 | 0.244 | 0.266 | 0.574 | 0.476 | 0.521 |
| | 3 | 0.563 | 0.468 | 0.511 | 0.574 | 0.476 | 0.521 |
| | 4 | 0.294 | 0.244 | 0.267 | 0.574 | 0.476 | 0.521 |
| | 5 | **0.574** | **0.477** | **0.521** | 0.574 | 0.476 | 0.521 |

### *6.3 Results*

Given that the pilot tasks were optional, we only received answers from a single team that submitted up to 5 different systems only for the NLU task considering the tourist and guide users. Table 5 shows the results extracted for this team. A baseline was not available for this task.

In past DSTCs, the evaluation was done by having teams submit a file with tracker output. In DSTC4, evaluations were conducted by having teams provide trackers as a web service. However, occasionally the web connection would time-out. For future evaluations, we suggest incorporating automatic reconnections when timeouts occur, and to add better handling of asynchronous communication data and packet-loss.

## 7 Conclusions

We have presented the official evaluation results of the Fourth Dialog State Tracking Challenge (DSTC4). This edition of the challenge has continued the tradition of its previous editions by providing a common testbed for the evaluation of Dialog State Tracking, one of the key tasks in Dialog Management. However, different from previous editions, which focused on human-machine dialogs, this edition has focused on dialog state tracking in human-human dialogs. The goal of the main task was to track dialog states for sub-dialog segments, which means that for each turn in a given sub-dialog, the tracker was required to fill out a frame of slot-value pairs considering all dialog history prior to the turn.

A total of seven teams participated in the main task with an overall number of twenty four entries submitted. Most of the submitted entries outperformed the provided baseline tracker system, which was based on a string matching strategy for identifying mentions of contents using the provided ontology as a reference. In a post-evaluation exercise of ensemble learning, results from multiple tracker submissions were combined. As a result, further significant improvements on dialog state tracking performance were observed.

In addition to the main task, this fourth edition of the challenge also proposed four pilot tasks with the objective of evaluating each of the core components needed for developing end-to-end dialog systems. More specifically, the proposed pilot tasks included: Spoken Language Understanding (SLU), Speech Act Prediction (SAP), Spoken Language Generation (SLG), and End-to-End System (EES). Only one team participated in the SLU pilot tasks with five different submissions for each of the two speaker roles involved in the provided datasets. This evaluation interestly showed that guide speech acts are significantly more predictable than tourist speech acts, while semantic tags are similarly predictable for both roles.

As final remarks, we would like to highlight that this challenge results have confirmed the feasibility of the state tracking task in human-human dialogs, which are much more unstructured and noisy than human-machine dialogs. We expect these

shared efforts on human dialog state tracking will contribute to progress in developing much more human-like dialog systems. Regarding the pilot task, on the other hand, we were able to test a new evaluation modality for dialog technology, which in our opinion constitutes a first step towards the effective development of distributed modular systems and a computational framework for collaborative end-to-end system evaluation.

As final recommendation, we suggest to continue pursuing the study of human-human dialogs as a means for better modeling and understanding the complexity of the pragmatic phenomena, as well as to include new languages to explore the feasibility of cross-language and/or multilingual approaches to dialog management. Similarly, we recommend to continue the efforts on pilot tasks for the next editions of the challenges, to continuing moving in the direction of distributed and modular end-to-end system construction and evaluation.

## References

1. Banchs, R.E., D'Haro, L.F., Li, H.: Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. Audio, Speech, and Language Processing, IEEE/ACM Transactions on **23**(3), 472–482 (2015)
2. Henderson, M., Thomson, B., Williams, J.: The second dialog state tracking challenge. In: 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, p. 263 (2014)
3. Henderson, M., Thomson, B., Williams, J.D.: The third dialog state tracking challenge. In: Spoken Language Technology Workshop (SLT), 2014 IEEE, pp. 324–329. IEEE (2014)
4. Kim, S., D'Haro, L.F., Banchs, R.E., Williams, J., Henderson, M.: Dialog state tracking challenge 4 handbook (2015). `http://www.colips.org/workshop/dstc4/Handbook_DSTC4.pdf`
5. Kim, S., D'Haro, L.F., Banchs, R.E., Williams, J., Henderson, M.: Dialog state tracking challenge 4 pilot task guidelines (2015). `http://www.colips.org/workshop/dstc4/DSTC4_pilot_tasks.pdf`
6. Lee, S., Eskenazi, M.: Recipe for building robust spoken dialog state trackers: Dialog state tracking challenge system description. In: Proceedings of the SIGDIAL 2013 Conference, pp. 414–422 (2013)
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311–318. Association for Computational Linguistics (2002)
8. Smith, R.W.: Comparative error analysis of dialog state tracking. In: 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, p. 300 (2014)
9. Williams, J., Raux, A., Ramachandran, D., Black, A.: The dialog state tracking challenge. In: Proceedings of the SIGDIAL 2013 Conference, pp. 404–413 (2013)
10. Yeh, A.: More accurate tests for the statistical significance of result differences. In: Proceedings of the 18th conference on Computational linguistics-Volume 2, pp. 947–953. Association for Computational Linguistics (2000)