Dialog State Tracking Challenge 5

Handbook v3.0

http://workshop.colips.org/dstc5/

Seokhwan Kim[1], Luis Fernando D'Haro[1], Rafael E. Banchs[1],
Matthew Henderson[2], Jason Williams[3], and Koichiro Yoshino[4]

[1] Institute for Infocomm Research (I2R - A*STAR), Singapore
[2] Google, USA
[3] Microsoft Research, USA
[4] Nara Institute of Science and Technology (NAIST), Japan

June 15, 2016

TABLE OF CONTENTS

1. Motivation

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 testbed for this task, the first Dialog State Tracking Challenge (DSTC) was initiated [1], and then two more challenges (DSTC 2&3) [2][3] had been organized keeping the aim at human-machine conversations. On the other hand, the fourth challenge (DSTC 4) which has been most recently completed [4] has shifted the target of state tracking to human-human dialogs. In the challenge, a dialog state was defined for each sub-dialog segment level as a frame structure filled with slot-value pairs representing the main subject of the segment. Then, trackers were required to fill out the frame considering all dialog history prior to each turn in a given segment.

The previous DSTCs have contributed to the spoken dialog research community by providing opportunities for sharing the resources, comparing results among the proposed algorithms, and improving the state-of-the-art. However, the impacts of the outcomes from the challenges could be restricted to English dialogs only, because all the resources including the corpora, ontologies, and databases were collected under monolingual settings in English.

In the fifth challenge, we introduce a cross-language dialog state tracking task addressing the problem of adaptation to a new language. The goal of this task is to build a tracker in the target language with given the existing resources in the source language and their translations generated automatically by machine translation technologies to the target language. In addition to this main task, we propose a series of pilot tracks for the core components in developing end-to-end dialog systems also in the same cross-language settings. We expect that these shared efforts on cross-language tasks would contribute to progress in improving the language portability of state-of-the-art monolingual technologies and reducing the costs for building resources from the scratch to develop dialog systems in a resource-poor target language.

This document is distributed as follows: section 2 gives detailed information about the challenge (i.e. how to register, the competition schedule, main and optional tasks). In section 3, the data used during the challenge is described. Examples of the dialogs included are also given. Section 4 describes the evaluation metrics and format for the main task submissions. Then, in section 5, description of several tools included with the data is provided. These tools are intended to allow participants to check the data, to have a baseline system that participants can modify or combine with their proposed systems. In section 6, the JSON data formats used for the annotations are described in detail. Finally, in section 7 several frequent questions are addressed regarding the data and participation in the challenge.

2. Participation

In this challenge, participants will be provided with labelled human-computer dialogs to develop dialog state tracking algorithms. Algorithms will then be evaluated on a common set of held-out dialogs, offline, to enable comparisons. As well as a corpus of labelled dialogs, participants will be given code that implements the evaluation measurements and a baseline tracker.

2.1. *Rules*

Participation is welcomed from any research team (academic, corporate, non profit, government). Members of the organizational committee and advisory committee are permitted to participate. In general, the identity of participants will not be published or made public. In written results, teams will be identified as team1, team2, etc. There are 2 exceptions to this: (1) the organizers will verbally indicate the identities of all teams at the conference/workshop chosen for communicating results; and (2) participants may identify their own team label (e.g. team5), in publications or presentations, if they desire, but may not identify the identities of other teams. On submission of results, teams will be required to fill in a questionnaire which gives some broad details about the approach they adopted.

In particular, there is interest in writing trackers that would be feasible for use in live systems. We provide a few recommendations for participants to follow so trackers may be directly comparable to others. These recommendations are not enforced, but will be addressed in the questionnaire at evaluation time.

- A tracker should be able to run fast enough to be used in a real-time dialog system
- A tracker should not make multiple runs over the test set before outputting its results
- A tracker should not use information from the future of a dialog to inform its output at a given turn

2.2. *Schedule*

Below we provide the proposed schedule for the evaluation as well as the conference submissions.

Shared Tasks

01 Apr 2016	Registration opens
14 Apr 2016	Labeled training and development data is released
18 Jul 2016	Registration closes
21 Jul 2016	Unlabeled test data is released
27 Jul 2016	Entry submission deadline
29 Jul 2016	Evaluation results are released

Task Papers

19 Aug 2016	Paper submission deadline
09 Sep 2016	Paper acceptance notifications
16 Sep 2016	Camera-ready submission deadline
13-16 Dec 2016	Workshop is held @ SLT 2016

Announcements and discussion about the challenge will be conducted on the group mailing list. Participants should be on the mailing list. Instructions for joining can be found on the DSTC homepage and at section 8 in this document.

2.3. *Registration*

The procedure to register to the Fifth Dialog State Tracking Challenge is as follows:

- **STEP 1:** Complete the registration form (available online). Please allow for few days while your request is being processed.
- **STEP 2:** Download and print the End-User-License Agreement (EULA) for the TourSG dataset.
- **STEP 3:** Complete, sign and submit back the EULA: the scanned version to kims@i2r.a-star.edu.sg and the original form to the following address:

  Mr Teo Poh Heng
  Exploit Technologies Pte Ltd
  1 Fusionopolis Way, #19 -10 Connexis North Tower
  Singapore 138632
  Tel: +65-6478 8420

- **STEP 4:** You will receive a one-time password for downloading the dataset.

2.4. *Tasks*

2.4.1. *Main task*

The goal of the main task of the challenge is to track dialog states for sub-dialog segments. For each turn in a given sub-dialog, the tracker should fill out a frame of slot-value pairs considering all dialog history prior to the turn. The performance of a tracker will be evaluated by comparing its outputs with reference annotations.

In the development phase, participants will be provided with a training set of English dialogs and a development set of Chinese dialogs with manual annotations over frame structures. In the test phase, each tracker will be evaluated on the results generated for a test set of unlabeled Chinese dialogs. A baseline system and evaluation script will be provided along with the training data.

2.4.2. *Pilot tasks*

Four pilot tasks are available in DSTC5. They will follow the same cross-language approach as in the main task. Systems are to be trained on English dialogs and a small subset of Chinese dialogs will be provided for development purposes. Evaluations will be conducted over Chinese dialogs.

- **Spoken language understanding:** Tag a given utterance with speech acts and semantic slots.
- **Speech act prediction:** Predict the speech act of the next turn imitating the policy of one speaker.
- **Spoken language generation:** Generate a response utterance for one of the participants.
- **End-to-end system:** Develop an end-to-end system playing the part of a guide or a tourist. This task will be conducted only if at least one team and/or individual registers for each of the pilot tasks above.

2.4.3. *Open track*

DSTC5 registered teams and/or individuals are free to work and report results on any proposed task of their interest over the provided dataset.

3. Data

3.1. *General Characteristics*

In this challenge, participants will use TourSG corpus to develop the components. TourSG consists of dialog sessions on touristic information for Singapore collected from Skype calls between tour guides and tourists. All the recorded dialogs with the total length of 21 hours, for English and Chinese each, have been manually transcribed and annotated with speech act and semantic labels for each turn level.

Since each subject in these dialogs tends to be expressed not just in a single turn, but through a series of multiple turns, dialog states are defined in these conversations for each sub-dialog level. A full dialog session is divided into sub-dialogs considering their topical coherence and then they are categorized by topics. Each sub-dialog assigned to one of major topic categories will have an additional frame structure with slot value pairs to represent some more details about the subject discussed within the sub-dialog. (See examples of reference annotations in Section 3.2).

Different from DSTC4, in which English dialogs were used, DSTC5 will focus on developing and evaluating cross-language strategies for Dialog State Tracking. In this edition of the challenge, trackers must be trained on English dialogs but will be evaluated on Chinese dialogs. A small annotated development set of Chinese dialogs will be provided too.

Regarding pilot tasks, the same cross-language scenario will be performed. English dialogs will be used as training data and the systems will be evaluated on Chinese data. A small subset of Chinese dialogs will be provided along with the training data. For pilot tasks, annotations are provided at the utterance level and, accordingly, systems must deal with information at the utterance level. These annotations involve semantic slots, speech acts and full utterances.

In addition to the original dialogs, their translations generated by a machine translation (MT) system from English to Chinese for the training set and from Chinese to English for the development and test sets will be also given along with the word alignment information, so that participants will not need to run their own system to generate the translated pairs of the dialogs.

3.2. *Example of Dialog Annotations for the Main Task*

Speaker	Transcription	Annotation
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.	**Topic:** Accommodation **Tourist_ACT:** REQ **Guide_ACT**: ACK **Type:** Hostel
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.	
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.	

Table 1. Example of a transcription and annotation for a sub-dialog segment #1

Speaker	Transcription	Annotation
Guide	Let's try this one, okay?	**Topic:** Accommodation **GuideAct:** RECOMMEND **TouristAct:** ACK **INFO:** Pricerange **Name:** InnCrowd Backpackers Hostel
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.	
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.	

Table 2. Example of a transcription and annotation for a sub-dialog segment #2

3.3. *Example of Dialog Annotations for the Pilot Tasks*

Speaker	Semantic Tagged Utterance	Speech Act (Attribute)
Tourist	Can you give me some uh- tell me some <**DET CAT="PRICE"**> cheap rate <**/DET**> <**LOC FROM-TO="NONE" REL="NONE" CAT="HOTEL"**> hotels <**/LOC**>, because I'm planning just to leave my bags there and go somewhere take some pictures.	QST (RECOMMEND) INI (EXPLAIN)
Guide	Okay. I'm going to recommend firstly you want to have a <**DET CAT="MAIN"**> backpack type <**/DET**> of <**LOC FROM-TO="NONE" REL="NONE" CAT="HOTEL"**> hotel <**/LOC**>, right?	FOL (ACK) INI (RECOMMEND) QST (PREFERENCE)
Tourist	Yes. I'm just gonna bring my backpack and my buddy with me. So I'm kinda looking for a hotel that is <**DET CAT="PRICE"**> not that expensive <**/DET**>. Just gonna leave our things <**LOC FROM-TO="NONE" REL="NONE" CAT="HOTEL"**> there <**/LOC**> and, you know, stay out the whole day.	RES (POSITIVE) RES (PREFERENCE| EXPLAIN)
Guide	Okay. Let me get you hm hm. So you don't mind if it's a bit uh <**DET CAT="MAIN"**> not so roomy <**/DET**> like hotel because you just back to sleep.	FOL (ACK) QST (PREFERENCE)
Tourist	Yes. Yes. As we just gonna put our things <**LOC FROM-TO="NONE" REL="NONE" CAT="HOTEL"**> there <**/LOC**> and then go out to take some pictures.	RES (POSITIVE) RES (PREFERENCE| EXPLAIN)
Guide	Okay, um-	FOL (ACK)
Tourist	Hm.	-

Table 3. Example of utterance level annotations for sub-dialog segment #1

Speaker	Transcription	Annotation
Guide	Let's try <LOC FROM-TO="NONE" REL="NONE" CAT="HOTEL">this one</LOC>, okay?	INI (RECOMMEND)
Tourist	Okay.	FOL (ACK)
Guide	It's <LOC FROM-TO="NONE" REL="NONE" CAT="HOTEL"> InnCrowd Backpackers Hostel </LOC> in <AREA FROM-TO="NONE" REL="NONE" CAT="CITY"> Singapore </AREA>. If you take <DET CAT="MAIN"> a dorm bed </DET> per person only <FEE CAT="MAIN"> twenty dollars </FEE>. If you take <DET CAT="MAIN"> a room </DET>, it's <DET CAT="MAIN"> two single beds </DET> at <FEE CAT="MAIN"> fifty nine dollars </FEE>.	INI (RECOMMEND) FOL (INFO|HOW_MUCH)
Tourist	Um. Wow, that's good.	FOL (POSITIVE)
Guide	Yah, <FEE CAT="MAIN"> prices </FEE> are based on <DET CAT="MAIN"> per person per bed or dorm </DET>. But this one is room. So it should be <FEE CAT="MAIN"> fifty nine </FEE> for the <DET CAT="MAIN"> two room </DET>. So you're actually paying about <FEE CAT="MAIN"> ten dollars </FEE> more <DET CAT="MAIN"> per person </DET> only.	FOL (INFO|HOW_MUCH)
Tourist	Oh okay. That's- <FEE CAT="MAIN"> the price </FEE> is reasonable actually. It's good.	FOL (ACK) FOL (POSITIVE)

Table 4. Example of utterance level annotations for a sub-dialog segment #2

3.4. *Data Available for DSTC5*

For the purposes of the DSTC5 Challenge, the TourSG corpus has been divided in the following three parts:

1. **Training set:** manual transcriptions and annotations at both utterance and sub-dialog levels will be provided for 35 English dialogs (the whole set for DSTC4) for training the trackers. For every utterance, 5-best results generated by a MT system from English to Chinese are also given along with the word alignment information.
2. **Development set:** similar to the training set, but for 2 Chinese dialogs with 5-best MT results in English also with word alignment information.
3. **Test set:** for the main task, manual transcriptions and their aligned MT results will be provided for 12 Chinese dialogs for evaluating the trackers. In the case of the pilot tasks, 5 Chinese dialogs will be used for evaluating each task.

The three datasets will be released free of charge to all registered challenge participants after signing a license agreement with ETPL-A*STAR. The dataset will include transcribed and annotated dialogs, as well as ontology objects describing the annotations.

4. Evaluation

4.1. *Main Task*

A system for the main task should generate the tracking output for every utterance in a given log file described in Section 6.1. 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 in 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 challenge, the following two sets of evaluation metrics are used for the main task:

- 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 that are correctly filled
  - Recall: Fraction of slot-value pairs in the gold standard labels that are 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 in whole frame-level, the others can show the partial correctness in each slot-value level.

4.2. *Pilot Tasks*

As the TourSG corpus constitutes a collection of conversations between two specific roles: a tour guide and a tourist, pilot tasks are to be focalized in modeling one of the two interlocutor roles of the TourSG dataset. In this sense, each pilot task has two primary subtasks, one related to the modeling of the tour guide and the other related to the modeling of the tourist. Each subtask can be better defined in terms of the input data the system should use and the output data it should produce.

4.2.1. *Spoken Language Understanding (SLU)*

In SLU task, the input to the systems will be the utterances from both the tourist and the guide, and the system must produce both semantic tags and speech acts only for the utterances spoken by either a tourist or a tour guide. The following evaluation metrics are used for the SLU task:

- Semantic tags
  - 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 of precision and recall
- Speech acts
  - 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 of precision and recall

4.2.2. *Speech Act Prediction (SAP)*

In SAP task, the input to the systems will be the utterances and annotations (semantic tags and speech acts) from the other speaker along with the resulting semantic tags for the next utterances by the target speaker, and the system must produce the speech acts for the corresponding utterances. The following evaluation metrics are used for the SAP task:

- Speech acts
  - 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 of precision and recall

4.2.3. *Spoken Language Generation (SLG)*

In SLG task, the input to the systems will be the semantic tags and speech acts from the target speaker only, and the system must produce the final surface form for the corresponding utterances. The following similarity metrics are used for the SLG task:

- BLEU: Geometric average of n-gram precision (for n = 1, 2, 3, 4) of the system generated utterance with respect to reference utterance.
- 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.

5. Included Scripts and Tools

As in the previous DSTCs, the DSTC 5 evaluation includes a set of useful scripts and tools[1] for dealing with the provided data. Below a brief description of the available tools is provided.

5.1. *Main task*

A simple baseline tracker for the main task is provided with the datasets. It is originally based on the one used for 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 segment to the current turn. To adapt it for the cross-language execution, the following two different methods are implemented in the tracker.

- Method 1: The translated utterances from Chinese to English are matched to the English entries in the original ontology.
- Method 2: The Chinese utterances are matched to the translated entries in the ontology from English to Chinese.

For both methods, only the top-1 hypothesis of the 5-best translations is used for each matching. 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 doesn't consider any semantic or discourse aspects from given dialogs, its performance is very limited and there is much room for improvement. The source code for the baseline tracker is included in *baseline.py*, please look there for full details on the implementations.

For running the baseline tracker, FuzzyWuzzy[2] package should be installed. And you should have a scripts directory with a config directory within it. The config directory contains the definitions of the datasets, e.g. *dstc5_dev.flist* which enumerates the sessions in the development set of DSTC 5. It also contains the ontology objects in *ontology_dstc5.json*. You can run the baseline tracker like so:

```
python scripts/baseline.py --dataset dstc5_dev --dataroot data --trackfile
baseline_dev.json --ontology scripts/config/ontology_dstc5.json --method 1
```

This will create a file *baseline_dev.json* with a tracker output object. The structure and contents of the output can be checked using *check_main.py*:

[1] https://github.com/seokhwankim/dstc5
[2] https://pypi.python.org/pypi/fuzzywuzzy

```
python scripts/check_main.py --dataset dstc5_dev --dataroot data --ontology scripts/config/ontology_dstc5.json --trackfile baseline_dev.json
```

This should output `Found no errors, trackfile is valid'. The checker is particularly useful for checking the tracker output on an unlabelled test set, before submitting it for evaluation in the challenge.

The evaluation script, *score_main.py* can be run on the tracker output like so:

```
python scripts/score_main.py --dataset dstc5_dev --dataroot data --trackfile baseline_dev.json --scorefile baseline_dev.score.csv --ontology scripts/config/ontology_dstc5.json
```

This creates a file *baseline_dev.score.csv* which lists all the metrics and we can use *report_main.py* to format these results:

```
python scripts/report_main.py --scorefile baseline_dev.score.csv
```

5.2. *Pilot task: SLU*

A simple baseline system for the cross-language SLU pilot task is also provided. It trains a pair of SVM and CRF models for multi-label speech act prediction and semantic tagging, respectively, from the English training dataset. Then, the trained models are used in analyzing the English translations of each Chinese utterance in the test set. Finally, the predicted labels are projected into the original utterances in Chinese through the word alignment information.

For running the baseline system, NLTK[3] and Scikit-learn[4] should be pre-installed. Then, you can run the system with the following commands:

```
python scripts/baseline_slu.py --trainset dstc5_train --testset dstc5_dev --dataroot data --roletype TOURIST --modelfile slu.baseline.tourist --outfile slu.baseline.tourist.json
```

The structure and contents in the generated output file *slu.baseline.tourist.json* can be checked using *check_slu.py*:

```
python scripts/check_slu.py --dataset dstc5_dev --dataroot data/ --jsonfile slu.baseline.tourist.json --ontology scripts/config/ontology_dstc5.json --roletype TOURIST
```

Once the checker outputs `Found no errors, trackfile is valid', the evaluation script *score_slu.py* can be run on the system output like so:

```
python scripts/score_slu.py --dataset dstc5_dev --dataroot data/ --jsonfile slu.baseline.tourist.json --ontology scripts/config/ontology_dstc5.json --roletype TOURIST --scorefile slu.baseline. tourist.score.csv
```

This creates a file *slu.baseline.tourist.score.csv* with the scores for all the metrics.

[3] http://www.nltk.org/
[4] http://scikit-learn.org/

5.3. *Pilot task: SAP*

The provided baseline for SAP task trains SVM models for multi-label speech act prediction using the 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 the features are language-independent, the model trained on the English training set can be also used for SAP on the Chinese test set.

For SAP pilot task, the format of the training and development data should be converted according to the required information in both input and output files with the following commands:

```
python scripts/convert_sap.py --dataset dstc5_train --dataroot data
python scripts/convert_sap.py --dataset dstc5_dev --dataroot data
```

After these conversions, the following four files will be generated for each session: *sap.guide.in.json*, *sap.guide.label.json*, *sap.tourist.in.json*, and *sap.tourist.label.json*. In the evaluation phase, only **.in.json* files will be provided for the test set to predict the information that should be included in **.label.json* files.

For running the baseline system, Scikit-learn should be pre-installed. Then, you can run the system with the following commands:

```
python scripts/baseline_sap.py --trainset dstc5_train --testset dstc5_dev --
dataroot data --roletype TOURIST --modelfile sap.baseline.tourist.model --
outfile sap.baseline.tourist.json
```

The structure and contents in the generated output file *sap.baseline.tourist.json* can be checked using *check_sap.py*:

```
python scripts/check_sap.py --dataset dstc5_dev --dataroot data/ --jsonfile
sap.baseline.tourist.json --ontology scripts/config/ontology_dstc5.json --
roletype TOURIST
```

Once the checker outputs `Found no errors, trackfile is valid', the evaluation script *score_sap.py* can be run on the system output like so:

```
python scripts/score_sap.py --dataset dstc5_dev --dataroot data/ --jsonfile
sap.baseline.tourist.json --ontology scripts/config/ontology_dstc5.json --
roletype TOURIST --scorefile sap.baseline.tourist.score.csv
```

This creates a file *sap.baseline.tourist.score.csv* with the scores for all the metrics.

5.4. *Pilot task: SLG*

The SLG baseline is based on an example-based language generation approach using *k*-nearest neighbors algorithm on the vector space with the speech act and semantic tags features. For each input, the system finds the most similar instance in the English training set, and then outputs the top-1 hypothesis of its Chinese translations as the generated result.

Similar to SAP task, first of all, the file conversion should be done with the following commands:

```
python scripts/convert_slg.py --dataset dstc5_train --dataroot data
python scripts/convert_slg.py --dataset dstc5_dev --dataroot data
```

The following four files should be located in the directory for each session: *slg.guide.in.json*, *slg.guide.label.json*, *slg.tourist.in.json*, and *slg.tourist.label.json*. In the evaluation phase, only **.in.json* files will be provided for the test set to predict the information that should be included in **.label.json* files.

For running the SLG baseline, Scikit-learn and Numpy should be pre-installed. Then, you can run the system with the following commands:

```
python scripts/baseline_slg.py --trainset dstc5_train --testset dstc5_dev --dataroot data --roletype TOURIST --outfile slg.baseline.tourist.json
```

The structure and contents in the generated output file *slg.baseline.tourist.json* can be checked using *check_slg.py*:

```
python scripts/check_slg.py --dataset dstc5_dev --dataroot data/ --jsonfile slg.baseline.tourist.json --roletype TOURIST
```

Once the checker outputs `Found no errors, trackfile is valid', the evaluation script *score_slg.py* can be run on the system output like so:

```
python scripts/score_slg.py --dataset dstc5_dev --dataroot data/ --jsonfile slg.baseline.tourist.json --scorefile slg.baseline.tourist.score.csv --roletype TOURIST
```

This creates a file *slg.baseline.tourist.score.csv* with the scores for all the metrics.

5.5. *Other Tools*

There are a few other scripts included which may be of use for participants:

- **dataset_walker.py:** A Python script which makes it easy to iterate through a dataset specified by file list (.flist) in scripts/config. When the script is called without arguments it outputs the content of all the training data on the terminal.
- **ontology_reader.py**: A Python script which makes it easy to get the information from the ontology.

6. JSON Data Formats

The datasets are distributed as collections of dialogs, where each dialog has a log.json file containing a Log object in JSON, and possibly a label.json containing a Label object in JSON representing the annotations. Every session also has a translations.json that consists of the translated utterances from Chinese to English or English to Chinese generated by a MT system. Also distributed with the data is an Ontology JSON object, which describes the ontology/domain of the sessions. The below sections describe the structure of the Log, Label, Translations and Ontology objects.

6.1. *Log Objects*

The *log.json* file includes the information for each session between a given tourist and a given guide. The JSON files were generated following below the specification:

- session_id: a unique ID for this session (integer)
- session_date: the date of the call, in yyyy-mm-dd format (string)
- session_time: the time the call was started, in hh:mm:ss format (string)
- lang: the language used in the session (string: “en”/”cn")

- guide_id: a unique ID for the guide participated in this session (string)
- tourist_id: a unique ID for the tourist participated in this session (string)
- tourist_age: the age of the tourist (integer)
- tourist_sex: the gender of the tourist (string: “F”/“M”)
- tourist_visited_sg: whether the tourist has visited or not Singapore in the past (string: “Y”/“N”)
- utterances: [
    - utter_index: the index of this utterances in the session starting at 0 (integer)
    - speaker: the speaker of this utterance (string: “GUIDE”/“TOURIST”)
    - transcript: the transcribed text of this utterance (string). Filler disfluencies in the recorded utterance are annotated with preceding percent sign (%) like “%ah”, “%eh”, “%uh”, or “%um”.
    - segment_info: [
        - topic: the topic category of the dialog segment that this utterance belongs to (string: “OPENING” / “CLOSING” / “ITINERARY” / “ACCOMMODATION” / “ATTRACTION” / “FOOD” / “SHOPPING” / “TRANSPORTATION”)
        - target_bio: the indicator with BIO scheme whether this utterance belongs to a segment considered as a target for the main task or not. The value for this key should be ‘B’ if this utterance is located at the beginning of a target session or ‘I’ if the utterance is not at the beginning but inside the target session. Otherwise, it is assigned to ‘O’. (string: “B”/“I”/“O”)
        - guide_act: the dialog act of the guide through this segment (string: “QST” / “ANS” / “REQ” / “REQ_ALT” / “EXPLAIN” / “RECOMMEND” / “ACK” / “NONE”)
        - tourist_act: the dialog act of the tourist through this segment (string: “QST” / “ANS” / “REQ” / “REQ_ALT” / “EXPLAIN” / “RECOMMEND” / “ACK” / “NONE”)
        - initiativity: whether this segment is initiated by the guide or the tourist (string: “GUIDE” / “TOURIST”)

    ]

]

6.2. *Label Objects*

The annotations for each segment are given in the *label.json* file. The json object in the label file consists of three different types of labels: frame structures for the main task and speech acts and semantics for the other pilot tasks. Below is the specification of the object:

- session_id: a unique ID for this session (integer)
- utterances: [
    - utter_index: a unique ID for this session (integer)
    - frame_label
        - SLOT: [list of values (string)]
    - speech_act: [
        - act: speech act category (string)
        - attributes: [list of attributes (string)]

    ]

    - semantic_tagged: [list of tagged utterances (string)]

]

6.2.1. *Frame labels*

The gold standard frame structure for the dialog segment that the current utterance belongs to is given as the object value of the ‘frame_label’ key. Each object consists of a set of a slot and a list of values pairs defined for the topic category of a given segment. Slots can be categorized into two different types: regular slots and ‘INFO’ slot. Each regular slot represents a major subject defined for a given topic and it should be filled with particular values mainly discussed in the current segment. Below is the list of regular slots for every topic category and their descriptions.

- ACCOMMODATION
  - PLACE: It refers to the names of accommodations discussed in a given segment
  - TYPE_OF_PLACE: It refers to the types of accommodations discussed in a given segment
  - NEIGHBOURHOOD: It refers to the geographic areas where the accommodations are located
- ATTRACTION
  - PLACE: It refers to the names of attractions discussed in a given segment
  - TYPE_OF_PLACE: It refers to the types of attractions discussed in a given segment
  - NEIGHBOURHOOD: It refers to the geographic areas where the attractions are located
  - ACTIVITY: It refers to the touristic activities discussed in a given segment
  - TIME: It refers to the discussed time slots to visit the attractions
- FOOD
  - PLACE: It refers to the names of places for eating discussed in a given segment
  - TYPE_OF_PLACE: It refers to the types of places for eating discussed in a given segment
  - NEIGHBOURHOOD: It refers to the geographic areas where the eating places are located
  - CUISINE: It refers to the cuisine types discussed in a given segment
  - DISH: It refers to the names of dishes discussed in a given segment
  - DRINK: It refers to the names of drinks discussed in a given segment
  - MEAL_TIME: It refers to the discussed time slots for eating
- SHOPPING
  - PLACE: It refers to the names of places for shopping discussed in a given segment
  - TYPE_OF_PLACE: It refers to the types of places for shopping discussed in a given segment
  - NEIGHBOURHOOD: It refers to the geographic areas where the shopping places are located
  - TIME: It refers to the discussed time slots for shopping
- TRANSPORTATION
  - TYPE: It refers to the types of transportation discussed in a given segment
  - TO: It refers to the destinations discussed in a given segment
  - FROM: It refers to the origins discussed in a given segment
  - LINE: It refers to the MRT lines discussed in a given segment
  - STATION: It refers to the train stations discussed in a given segment
  - TICKET: It refers to the types of tickets for transportation

In addition to the regular slots, a frame could have a special slot named ‘INFO’ to indicate the subjects that are discussed in a given segment but not directly related to any particular values of other slots. For example, 'INFO' slot in a frame for ‘FOOD’ topic could be filled in with 'DISH' value if the segment deals with some

general contents regarding dishes. But, when the speakers are talking about a specific dish, the frame has the corresponding value for the 'DISH' slot instead of a 'DISH' value for 'INFO' slot. Below is the list of 'INFO' slot values for each topic category and the descriptions about the target contents to be annotated with them.

ACCOMMODATION	
Amenity	amenities of accommodations
Architecture	architectural aspects of accommodations
Booking	booking for accommodations
Check-in	checking in for accommodations
Check-out	checking out of accommodations
Cleanness	cleanness of accommodations
Facility	facilities of accommodations
History	history of accommodations
Hotel rating	rating of accommodations
Image	dialogs with showing some images of accommodations
Itinerary	itinerary focusing on accommodations
Location	locations of accommodations
Map	dialogs with showing maps of the areas near accommodations
Meal included	meal plans provided by accommodations
Name	names of accommodations
Preference	tourists' preferences in looking for accommodations
Pricerange	room charges for accommodations
Promotion	discount promotions for accommodations
Restriction	any restrictions in accommodations
Room size	room sizes in accommodations
Room type	room types in accommodations
Safety	safety issues in accommodations

ATTRACTION	
Activity	tourist activities
Architecture	architectural aspects of tourist attractions
Atmosphere	atmosphere of tourist attractions
Audio guide	audio guide provided by tourist attractions
Booking	booking for tourist attractions
Dresscode	dress code for tourist attractions
Duration	time durations for visiting tourist attractions
Exhibit	exhibits shown in tourist attractions
Facility	facilities of tourist attractions
Fee	admission charges for tourist attractions
History	history of tourist attractions
Image	dialogs with showing some images of tourist attractions
Itinerary	itinerary focusing on tourist attractions

Location	locations of tourist attractions
Map	dialogs with showing maps of the areas near tourist attractions
Name	names of tourist attractions
Opening hour	operation hours of tourist attractions
Package	package tours or tickets for tourist attractions
Place	general discussion about tourist places without specifying target attractions
Preference	tourists' preferences in visiting tourist attractions
Promotion	discount promotions for tourist attractions
Restriction	any restrictions in tourist attractions
Safety	safety issues in tourist attractions
Schedule	schedules for exhibitions or shows in tourist attractions
Seat	seat information for shows in tourist attractions
Ticketing	ticketing information for tourist attractions
Tour guide	guided tour for tourist attractions
Type	types of tourist attractions
Video	dialogs with showing some video clips of tourist attractions
Website	dialogs with showing websites of tourist attractions

FOOD	
Cuisine	cuisine type for foods
Delivery	delivery services of foods
Dish	general discussion about dishes without specifying targets
History	history of foods or restaurants
Image	dialogs with showing some images of foods or restaurants
Ingredient	ingredients for foods
Itinerary	itinerary focusing on dining
Location	locations of restaurants
Opening hour	operation hours of restaurants
Place	general discussion about dining places without specifying targets
Preference	tourists' preferences for dining
Pricerange	price ranges for dining
Promotion	discount promotions for dining
Restriction	any restrictions in dining
Spiciness	spiciness about foods
Type of place	types of food places

SHOPPING	
Brand	brands of goods
Duration	time durations for shopping
Image	dialogs with showing some images of goods or shopping places
Item	shopping items
Itinerary	itinerary focusing on shopping
Location	locations of shopping places

Map	dialogs with showing maps of the areas near shopping places
Name	names of shopping places
Opening hour	operation hours of shopping places
Payment	payment options available at shopping places
Place	general discussion about shopping places without specifying targets
Preference	tourists' preferences for shopping
Pricerange	price ranges for shopping
Promotion	discount promotions for shopping
Tax refund	information about tax refund for tourists
Type	types of shopping places

TRANSPORTATION	
Deposit	information about deposits in tickets
Distance	traveling distance between origin and destination
Duration	travel time between origin and destination
Fare	transportation expenses
Itinerary	itinerary focusing on local transportation
Location	locations of train stations, bus stops, or terminals
Map	dialogs with showing train or bus route maps
Name	names of train stations, bus stops, or terminals
Preference	tourists' preferences in travelling with local transportations
Schedule	schedules for public transportations
Service	services related to transportations
Ticketing	ticketing information for local public transportations
Transfer	information about transit transfer to another line or another type of transportation
Type	types of transportation

The set of candidate values for each slot can be found in the ontology object (Section 6.4). Since each slot has a list of string values in its JSON object, multiple values can be assigned to a single slot, if more than one subject regarding the particular slot type are discussed in a given segment. All the annotations have been done considering not only the occurrences of relevant words for each candidate value in the surface of the segment, but also the correlations with the main subject of conversation at the moment that the sub-dialog was going on.

6.2.2. *Speech acts*

Since the speech acts were originally analyzed for each sub-utterance unit divided based on the pauses in the recordings and then combined into the full utterance level, each utterance could have more than one speech act objects if it was generated by concatenating its multiple sub-utterances. Thus, a list of speech act annotations is taken as the value for the 'speech_act' key of a given utterance.

Each object has two types of information: speech act category and attributes. Every sub-utterance should belong to one of the four basic speech act categories that denote the general role of the utterance in the current

dialog flow. More specific speech act information can be annotated by combination with attributes. By contrast to act category, there's no constraint on the number of attributes for a single utterance. Thus, a sub-utterance can have no attribute or more than one attributes in the list object. Below are the list of speech act categories and attributes with their descriptions.

- Speech act categories
    - QST (QUESTION) used to identify utterances that pose either a question or a request
    - RES (RESPONSE) used to identify utterances that answer to a previous question or a previous request
    - INI (INITIATIVE) used to identify utterances that constitute new initiative in the dialog, which does not constitute either a question, request, answer or follow up action to a previous utterance
    - FOL (FOLLOW) a response to a previous utterance that is not either a question or a request
- Speech act attributes
    - ACK: used to indicate acknowledgment, as well as common expressions used for grounding
    - CANCEL: used to indicate cancelation
    - CLOSING: used to indicate closing remarks
    - COMMIT: used to identify commitment
    - CONFIRM: used to indicate confirmation
    - ENOUGH: used to indicate/request that no more information is needed
    - EXPLAIN: used to indicate/request an explanation/justification of a previous stated idea
    - HOW_MUCH: used to indicate money or time amounts
    - HOW_TO: used to request/give specific instructions
    - INFO: used to indicate information request
    - NEGATIVE: used to indicate negative responses
    - OPENING: used to indicate, opening remarks
    - POSITIVE: used to indicate positive responses
    - PREFERENCE: used to indicate/request preferences
    - RECOMMEND: used to indicate/request recommendations
    - THANK: used to indicate thank you remarks
    - WHAT: used to indicate concept related utterances
    - WHEN: used to indicate time related utterances
    - WHERE used to indicate location related utterances
    - WHICH: used to indicate entity related utterances
    - WHO: used to indicate person related utterances and questions

6.2.3. *Semantic tags*

Similarly to speech acts, semantic tags were also annotated for each sub-utterance level. Thus it takes a list of tagged sub-utterances as its value, and the number of items in the list should be the same with the one for speech acts.

We defined below the main categories for semantic annotation:

- AREA: It refers to a geographic area but not a specific spot or location
- DET: It refers to user's criteria used or reasons why the user would like to decide spot.

- FEE: It refers to admission fees, price of services or any other fare.
- FOOD: It refers to any type of food or drinks.
- LOC: It refers to specific touristic spots or commerce/services locations.
- TIME: It refers to time, terms, dates, etc.
- TRSP. It refers to expressions related to transportation and transportation services.
- WEATHER: It refers to any expression related to weather conditions.

Some of them include also subcategories, relative modifiers and from-to modifiers (Table 3).

MAIN	SUBCAT	REL	FROM-TO
AREA	COUNTRY, CITY, DISTRICT, NEIGHBORHOOD	NEAR, FAR, NEXT, OPPOSITE, NORTH, SOUTH, EAST, WEST	FROM, TO
DET	ACCESS, BELIEF, BUILDING, EVENT, PRICE, NATURE, HISTORY, MEAL, MONUMENT, STROLL, VIEW	-	-
FEE	ATTRACTION, SERVICES, PRODUCTS	-	-
FOOD	-	-	-
LOC	TEMPLE, RESTAURANT, SHOP, CULTURAL, GARDEN, ATTRACTION, HOTEL, WATERSIDE, EDUCATION, ROAD, AIRPORT	NEAR, FAR, NEXT, OPPOSITE, NORTH, SOUTH, EAST, WEST	FROM, TO
TIME	DATE, INTERVAL, START, END, OPEN, CLOSE	BEFORE, AFTER, AROUND	-
TRSP	STATION, TYPE	NEAR, FAR, NEXT, OPPOSITE, NORTH, SOUTH, EAST, WEST	FROM, TO
WEATHER	-	-	-

Table 3. List of Categories and Modifiers for Semantic Annotations

The semantic tags and their categories are indicated as follows:

- <MAIN CAT=”SUBCAT” REL=”REL” FROM-TO=”FROM_TO”> at the beginning of the identified word or compound
- </TAG> at the end of the identified word or compound

When either no specific subcategory exists for a given semantic tag or it is not possible to select among the available subcategories, the ‘CAT’ field is assigned to cat=”MAIN”.

6.3. *Translations Objects*

The translations generated by a MT system for each dialog session are given in the *translations.json* file. Below is the specification of the object:

- session_id: a unique ID for this session (integer)
- lang_src: the source language (string: “en”/”cn”)
- lang_tgt: the target language (string: “en”/”cn”)
- utterances: [
    - utter_index: a unique ID for this session (integer)
    - translated [
        - hyp: a hypothesis from the 5-best MT results (string)
        - align: [list of aligned word pairs]

      ]

]

6.4. *Tracker Output Objects*

Tracker outputs should be organized following below the JSON specification for each task.

6.4.1. *Main task*

- dataset: the name of the dataset over which the tracker has been run (string)
- wall_time: the time in seconds it took to run the tracker (float)
- sessions: a list of results corresponding to each session in the dataset [
    - session_id: the unique ID of this session (integer)
    - utterances: [
        - utter_index: a unique ID for this session (integer)
        - frame_label: the tracker output for the segment that this utterance belongs to. The expected format for this object is same as the ones in the reference label objects. (Section 6.2.1)

      ]

6.4.2. *Pilot task: SLU*

- dataset: the name of the dataset over which the tracker has been run (string)
- wall_time: the time in seconds it took to run the tracker (float)
- task_type: “SLU”
- role_type: “GUIDE”/”TOURIST”
- sessions: a list of results corresponding to each session in the dataset [
    - session_id: the unique ID of this session (integer)
    - utterances: [
        - utter_index: a unique ID for this session (integer)
        - semantic_tagged: tagged utterance (string)
        - speech_act: [
            - act: speech act category (string)
            - attributes: [list of attributes (string)]

          ]]

6.4.3. *Pilot task: SAP*

- dataset: the name of the dataset over which the tracker has been run (string)
- wall_time: the time in seconds it took to run the tracker (float)
- task_type: “SLU”
- role_type: “GUIDE”/”TOURIST”
- sessions: a list of results corresponding to each session in the dataset [
    - session_id: the unique ID of this session (integer)
    - utterances: [
        - utter_index: a unique ID for this session (integer)
        - speech_act: [
            - act: speech act category (string)
            - attributes: [list of attributes (string)]

            ]

    ]

6.4.4. *Pilot task: SLG*

- dataset: the name of the dataset over which the tracker has been run (string)
- wall_time: the time in seconds it took to run the tracker (float)
- task_type: “SLU”
- role_type: “GUIDE”/”TOURIST”
- sessions: a list of results corresponding to each session in the dataset [
    - session_id: the unique ID of this session (integer)
    - utterances: [
        - utter_index: a unique ID for this session (integer)
        - generated: generated utterance (string)

        ]

6.5. ***Ontology Object***

The ontology object in *ontology_dstc5.json* describes the definitions of the frame structures for the main task and some additional domain knowledges in the following format:

- tagsets
    - TOPIC
        - SLOT: [list of possible values]
- knowledge
    - MRT_LINE
        - CODE: string
        - NAME: string

        - COLOR: string
    - SHOPPING
        - NAME: string
        - TYPE_OF_PLACE: [list of shopping place types]
    - RESTAURANT
        - NAME: string
        - TYPE_OF_PLACE: [list of restaurant types]
        - NEIGHBOURHOOD: [list of neighbourhoods]
        - CUISINE: [list of cuisines]
        - PRICERANGE: integer (from 1 to 5)
    - FOOD
        - NAME: string
        - CUISINE: string
    - MRT_STATION
        - NAME: string
        - CODE: [list of codes]
        - NEIGHBOURHOOD: [list of neighbourhoods]
    - HOTEL
        - NAME: string
        - TYPE_OF_PLACE: [list of hotel types]
        - NEIGHBOURTHOOD: [list of neighbourhoods]
        - RATING: integer (from 1 to 5)
        - PRICERANGE: integer (from 1 to 5)
    - ATTRACTION
        - NAME: string
        - TYPE_OF_PLACE: [list of attraction types]
        - NEIGHBOURHOOD: [list of neighbourhoods]
    - ROAD
        - NAME: string
        - NEIGHBOURHOOD: [list of neighbourhoods]
    - NEIGHBOURHOOD
        - REGION: string
        - DISTRICT: string
        - SUBDISTRICT: string
- translations
    - NAME: [list of 5-best translations in Chinese]

The slots and their values in the frame labels in both label object and tracker output object should be chosen following the definitions in the 'tagset' object. For each target topic category, a set of slots in the frame structure are given in the object. And the valid values are also specified as the value of the slot in this object. Two different types of values can be described in this list: the first type is for the static values and the other type is for referring to some information in the 'knowledge' object. For example, the 'MEAL_TIME' slot in the 'FOOD' topic frame could take some values from a list of static values:

"MEAL_TIME": ["Breakfast", "Dinner", "Lunch"].

On the other hand, the 'DISH' slot in the same frame has below the object as its value:

"DISH": [{ "slot": "NAME", "source": "DISH", "type": "knowledge" }],

which means that the list of 'NAME' values of 'DISH' objects in 'knowledge' part should be considered as the candidate values for the 'DISH' slot in the frame.

7. Frequently Asked Questions (FAQ)

7.1. *Do I or my company need to pay a license fee for getting the TourSG dataset?*

No, the TourSG dataset will be provided under a free of charge end-user-license to all participants in the Fourth Dialog State Tracking Challenge.

7.2. *Can I get the TourSG dataset without participating in the Challenge?*

Yes, but no free of charge end-user-license is available to non-participants. You or your company will need to pay a license fee for getting the TourSG dataset without participating in the Fourth Dialog State Tracking Challenge.

7.3. *Is participation in the main task of the Challenge mandatory?*

Participation in the main task of the Challenge is not mandatory for teams participating in at least one of the four pilot tasks or open track.

7.4. *Are baseline systems and evaluation scripts going to be provided?*

A baseline system and evaluation scripts will be provided only for the main task of the Challenge. Baselines and evaluation protocols for pilot tasks and open track are to be agreed directly with participants on such tasks.

7.5. *Is participation in the pilot tasks and open track of the Challenge mandatory?*

Participation in the pilot tasks and open track of the Challenge is optional for teams already participating in the main task of the Challenge.

8. Subscription to mailing list

To join the mailing list, send an email to listserv@lists.research.microsoft.com with 'subscribe DSTC' in the body of the message (without quotes). Joining the list is encouraged for those with an interest in the challenge, and is a necessity for those participating.

Post to the list using the address: dstc@lists.research.microsoft.com.

9. Organizing Committees

Seokhwan Kim - I2R A*STAR
Luis F. D'Haro - I2R A*STAR
Rafael E Banchs - I2R A*STAR
Matthew Henderson - Google
Jason Williams - Microsoft Research
Koichiro Yoshino - NAIST

10. References

[1] Jason D Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. The dialog state tracking challenge. In In Proceedings 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France, 2013.

[2] Matthew Henderson, Blaise Thomson, and Jason Williams. The Second Dialog State Tracking Challenge. In Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, p. 263. 2014.

[3] Henderson, Matthew, Blaise Thomson, and Jason Williams. The Third Dialog State Tracking Challenge. In Proceedings of IEEE Spoken Language Technology (2014).

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