Dialog State Tracking Challenge: Information for prospective participants

Jason D. Williams
Alan Black
Deepak Ramachandran
Antoine Raux

IEEE SLT
December 2012
What is dialog state tracking?

• In a spoken dialog system, given dialog history up to $t$, predict the user’s goal at time $t$

• A simple baseline: choose the top ASR result, perhaps thresholded by a confidence score

• But it’s possible to do better...
Why a challenge task?

• Work over the past 10+ years has shown that it is possible to outperform the ASR 1-best using statistical techniques ...
  ... but in some cases rules still perform better

• Variety of techniques have been proposed ...
  ... but different research sites use their own systems, so there have been virtually no comparative evaluations – we need a common testbed
Fixed corpus of dialogs

• A fixed corpus, not an end-to-end evaluation
  – Lower barrier to entry
  – No need to develop ASR, TTS, etc.
  – Facilitates direct comparisons of dialog state tracking algorithms, independent of other modules
Limitations of a fixed corpus

1. Experiments on a fixed corpus may not predict performance in deployment
   – Develop tracker on training data drawn from a particular distribution
   – Deploy tracker into system
   – Tracker causes system to follow a different distribution
   – Problem: train/test mismatch
   – We explicitly create train/test mismatch in the challenge

2. Does not directly measure improvement in whole-dialog performance (e.g., task completion)
   – The ultimate quantity of interest are whole-dialog measures like task completion
   – However, measuring whole-dialog performance precludes evaluation on a corpus
## Challenge schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Milestone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 July - 3 August 2012</td>
<td>Announced at SigDial; comment period on labeling and evaluation metrics</td>
</tr>
<tr>
<td>31 August 2012</td>
<td>Evaluation metrics and labeling guide published; labeling begins</td>
</tr>
<tr>
<td>12 November 2012</td>
<td>Pre-release of train3 data and helper scripts, for comment</td>
</tr>
<tr>
<td>~ 10 December 2012</td>
<td>Training data available; challenge begins (14 weeks)</td>
</tr>
<tr>
<td>22 March 2013</td>
<td>Test data available; evaluation begins</td>
</tr>
<tr>
<td>29 March 2013</td>
<td>Test results due to organizers</td>
</tr>
<tr>
<td>5 April 2013</td>
<td>Notification of results to teams</td>
</tr>
<tr>
<td>3 May 2013</td>
<td>SigDial 2013 deadline (4 weeks)</td>
</tr>
<tr>
<td>23-24 August 2013</td>
<td>SigDial conference, Metz, France</td>
</tr>
</tbody>
</table>
Data: Spoken Dialog Challenge

- Hosted by Dialog Research Center at CMU in 2010 and 2011-2012
- Bus timetables in Pittsburgh
- Systems from 3 research groups: A, B, C
- Real users
- Multiple time periods and system requirements
- Non-trivial scale: over 20K dialogs
Example call

Welcome to the Let's Go bus information system. To get help at any time, just say Help or press zero.

#noise#
Leaving from COAL. Did I get that right?
no
What can I do for you?
61c leaving downtown
The 61C. Did I get that right?
yes
Right Leaving from DOWNTOWN. Did I get that right?
yes
Alright Where do you wanna go?
McKeesport
Going to MCKEESPORT. When would you like to travel?
now
The next bus. Just a minute. Let me check that for you. The next 61C leaves FIFTH AVENUE AT MARKET at 11 oh 7 p.m. and arrives at MCKEESPORT TRANSPORTATION CENTER AT BAY NUMBER 2 at 12 oh 5 a.m.. To get more information about buses related to this trip, you can say, when is the next bus, or, when is the previous bus. To ask about a different trip, you can say, start a new query. If you are finished, you can say goodbye.
5 “slots”

• Bus routes: ~100
• From location: ~1K – 10K
  – Split into “desc”, “neighborhood”, “monument”
• To location: ~1K – 10K
  – Split into “desc”, “neighborhood”, “monument”
• Dates: ~100
• Times: ~1K
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Calls</th>
<th>Time period</th>
<th>Transcribed?</th>
<th>Labeled?</th>
</tr>
</thead>
<tbody>
<tr>
<td>train1a</td>
<td>Group A</td>
<td>1761</td>
<td>September 2009</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>train1c</td>
<td>Group A</td>
<td>1765</td>
<td>August 2009</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>train1b *</td>
<td>Group A</td>
<td>14,545</td>
<td>16 Months (2008-2009)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>train2</td>
<td>Group A</td>
<td>678</td>
<td>Summer 2010</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>train3</td>
<td>Group B</td>
<td>779</td>
<td>Summer 2010</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>test1</td>
<td>Group A</td>
<td>765</td>
<td>Winter 2011-12</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>test2</td>
<td>Group A</td>
<td>983</td>
<td>Winter 2011-12</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>test3</td>
<td>Group B</td>
<td>1037</td>
<td>Winter 2011-12</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>test4</td>
<td>Group C</td>
<td>451</td>
<td>Summer 2010</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

* will be available approx Jan 1, 2013
What is provided?

• Parsed system log files, in an easily readable format
• Offline recognition result with NBest result, for systems which did not produce online NBest lists
• Utterance transcriptions (training set)
• User goal labels (training set)

• The scoring tool that will be used in the evaluation stage
• Bus timetable database
• Challenge handbook (transcription and labeling guides)

• For the very keen: Raw system log files and utterance audio are available from dialrc.org
Tour of data

[see challenge handbook]
Labels

• For each utterance, the label files include:
  – Transcription of the words spoken
  – Indication of the correctness of each SLU hypothesis
Tour of labels

[see challenge handbook]
Evaluation overview

- **Assumption 1**: User’s goal is fixed, except when they “start over”
- **Assumption 2**: Guessing a value that hasn’t been observed on an N-Best list would give trivial improvements in accuracy

- With these assumptions, tracker output is a list of the form
  - (observed SLU hyp, score)
Example tracker output (route slot)

Sys transcript:  Which bus route?
Sys dialog acts: request(route)

Sorry, which bus route?
sorry(), request(route)

SLU hyps:
- inform(route=61c) 0.1
- inform(route=28x) 0.3
- inform(route=61b) 0.1

Tracker output (route slot):
- inform(route=61c) 0.1
- inform(route=28x) 0.3
- inform(route=61b) 0.1
- none 0.5

- inform(route=56u) 0.1
- inform(route=61d) 0.6

- none 0.1
10 tracker output lists at each turn:

- At each turn $t$, the tracker outputs:
  - List of (route, score)
  - List of (from.desc, score)
  - List of (from.neighborhood, score)
  - List of (from.monument, score)
  - List of (to.desc, score)
  - List of (to.neighborhood, score)
  - List of (to.monument, score)
  - List of (day, score)
  - List of (time, score)
  - List of (route, from.*, to.*, day, time, score)
What metrics are measured?

• 1-best hypothesis accuracy
• Mean reciprocal rank (mrr)
• Average probability assigned to correct item (avgp)
• Score calibration (L2 norm)
• ROC performance
  – Equal error rate (EER)
  – Correct accept at a false accept rate of 5% (ca05)
  – Correct accept at a false accept rate of 10% (ca10)
  – Correct accept at a false accept rate of 20% (ca20)
**When are metrics measured?**

<table>
<thead>
<tr>
<th>schedule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>schedule1</td>
<td>Include all turns (regardless of dialog context)</td>
</tr>
<tr>
<td>schedule2</td>
<td>Include a turn for a given concept only if:</td>
</tr>
<tr>
<td></td>
<td>• Concept appears on the SLU N-Best list in that turn, OR</td>
</tr>
<tr>
<td></td>
<td>• The system’s action references that concept in that turn (e.g., an explicit or implicit confirmation)</td>
</tr>
<tr>
<td>schedule3</td>
<td>Include only the last turn of the dialog</td>
</tr>
</tbody>
</table>
## Datasets

<table>
<thead>
<tr>
<th>dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>train3.sessions</td>
<td>All calls in train3</td>
</tr>
<tr>
<td>train3.half1.sessions</td>
<td>First half of calls in train3</td>
</tr>
<tr>
<td>train3.half2.sessions</td>
<td>Second half of calls in train3</td>
</tr>
<tr>
<td></td>
<td>(encourage participants to report performance by training on half1 and testing on half2)</td>
</tr>
<tr>
<td>train3.call1.sessions</td>
<td>The first call (for testing)</td>
</tr>
</tbody>
</table>

All datasets are here: installpath/config
For the very keen

• You *can* re-run SLU (or ASR) if you want to, but...
  – You can’t guess a SLU hyp that’s not in the data
  – Please make it clear you’ve re-run ASR/SLU in your paper/system description
For the mischievous

• We’ve designed the challenge to have the low barriers to entry. We recognize it is possible for participants to exploit this design to overstate performance.

• Two obvious things not to do:
  – The tracker should *not* look ahead in the dialog
  – Don’t download the audio for the test data and label it
Example run with the baseline

> bin/baseline --dataset=train3.half2 \
  --dataroot=../data \
  --trackfile=track.json

The baseline is also a useful template for training and testing
What’s in a trackfile

[see challenge handbook]
Evaluating the baseline

> bin/score --dataset=train3.half2 \ 
  --dataroot=../data \ 
  --trackfile=track.json \ 
  --scorefile=score.csv
What’s in a score file

CSV with “slot, schedule, metric name, N utts, metric”

date,schedule1,accuracy,4459,0.891231217762
date,schedule1,avgp,4459,0.892024676833
date,schedule1,l2,4459,0.0797279581255
date,schedule1,mrr,4459,0.933393137475
date,schedule1,roc.ca05,4459,0.846602377215
date,schedule1,roc.ca10,4459,0.883606189729
date,schedule1,roc.ca20,4459,0.891231217762
date,schedule1,roc.eer,4459,0.0681767212379
date,schedule2,accuracy,189,0.820105820106
date,schedule2,avgp,189,0.660067010582
date,schedule2,l2,189,0.172862696576
date,schedule2,mrr,189,0.888888888889
date,schedule2,roc.ca05,189,0.470899470899
...

246 rows in total
Create a report

> bin/report --scorefile=score.csv

---

**schedule1**

<table>
<thead>
<tr>
<th>route</th>
<th>from.d</th>
<th>from.m</th>
<th>from.n</th>
<th>to.des</th>
<th>to.mon</th>
<th>to.nei</th>
<th>date</th>
<th>time</th>
<th>joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>4459</td>
<td>4459</td>
<td>4459</td>
<td>4459</td>
<td>4459</td>
<td>4459</td>
<td>4459</td>
<td>4459</td>
<td>4459</td>
</tr>
</tbody>
</table>

| accuracy | 0.7540 | 0.7899 | 1.0000 | 1.0000 | 0.8143 | 1.0000 | 1.0000 | 0.8912 | 0.9551 | 0.4532 |
| avgp    | 0.6686 | 0.7226 | 1.0000 | 1.0000 | 0.7840 | 1.0000 | 1.0000 | 0.8920 | 0.9401 | 0.3843 |
| l2      | 0.2341 | 0.1729 | 0.0000 | 0.0000 | 0.1454 | 0.0000 | 0.0000 | 0.0797 | 0.0441 | 0.5463 |
| mrr     | 0.7947 | 0.8595 | 1.0000 | 1.0000 | 0.8663 | 1.0000 | 1.0000 | 0.9334 | 0.9597 | 0.4741 |
| roc.ca05| 0.3292 | 0.4927 | 1.0000 | 1.0000 | 0.6152 | 1.0000 | 1.0000 | 0.8466 | 0.9551 | 0.1390 |
| roc.ca10| 0.4898 | 0.6185 | 1.0000 | 1.0000 | 0.7600 | 1.0000 | 1.0000 | 0.8836 | 0.9551 | 0.2166 |
| roc.ca20| 0.7392 | 0.7813 | 1.0000 | 1.0000 | 0.8143 | 1.0000 | 1.0000 | 0.8912 | 0.9551 | 0.2808 |
| roc.eer | 0.2523 | 0.2671 | 0.0000 | 0.0000 | 0.1698 | 0.0000 | 0.0000 | 0.0682 | 0.1070 | 0.3409 |

---

**basic stats**

- dataset: train3.half2
- scorer_version: 0.3
- sessions: 344
- total_wall_time: 2.72199988365
- turns: 4459
- wall_time_per_turn: 0.000610450747623
Where is ...

• Pointers to everything here:
  
  research.microsoft.com/events/dstc

• Handbook
• Training data (next week)
  – Two packages – one from MSR, one from Honda
• Helper scripts + baseline system
• Mailing list

• Test data (in March)
Thanks to ... our advisory board

- Daniel Boies, Microsoft, Canada
- Paul Crook, Microsoft, USA
- Maxine Eskenazi, Carnegie Mellon University, USA
- Milica Gasic, University of Cambridge, UK
- Dilek Hakkani-Tur, Microsoft, USA
- Helen Hastie, Heriot Watt University, UK
- Kee-Eung Kim, KAIST, Korea
- Ian Lane, Carnegie Mellon University, USA
- Sungjin Lee, Carnegie Mellon University, USA
- Teruhisa Misu, NICT, Japan
- Olivier Pietquin, SUPELEC, France
- Joelle Pineau, McGill University, Canada
- Blaise Thomson, University of Cambridge, UK
- David Traum, USC Institute for Creative Technologies, USA
- Luke Zettlemoyer, University of Washington, USA
Thanks to ... our sponsors

- Honda research institute
- Microsoft
Thanks to ... SigDial and DialRC

• The dialog state tracking challenge is endorsed by SigDial

• Raw data and labeling support provided by Dialog Research Center
Dialog State Tracking Challenge

research.microsoft.com/events/dstc

Jason D. Williams
Alan Black
Deepak Ramachandran
Antoine Raux