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

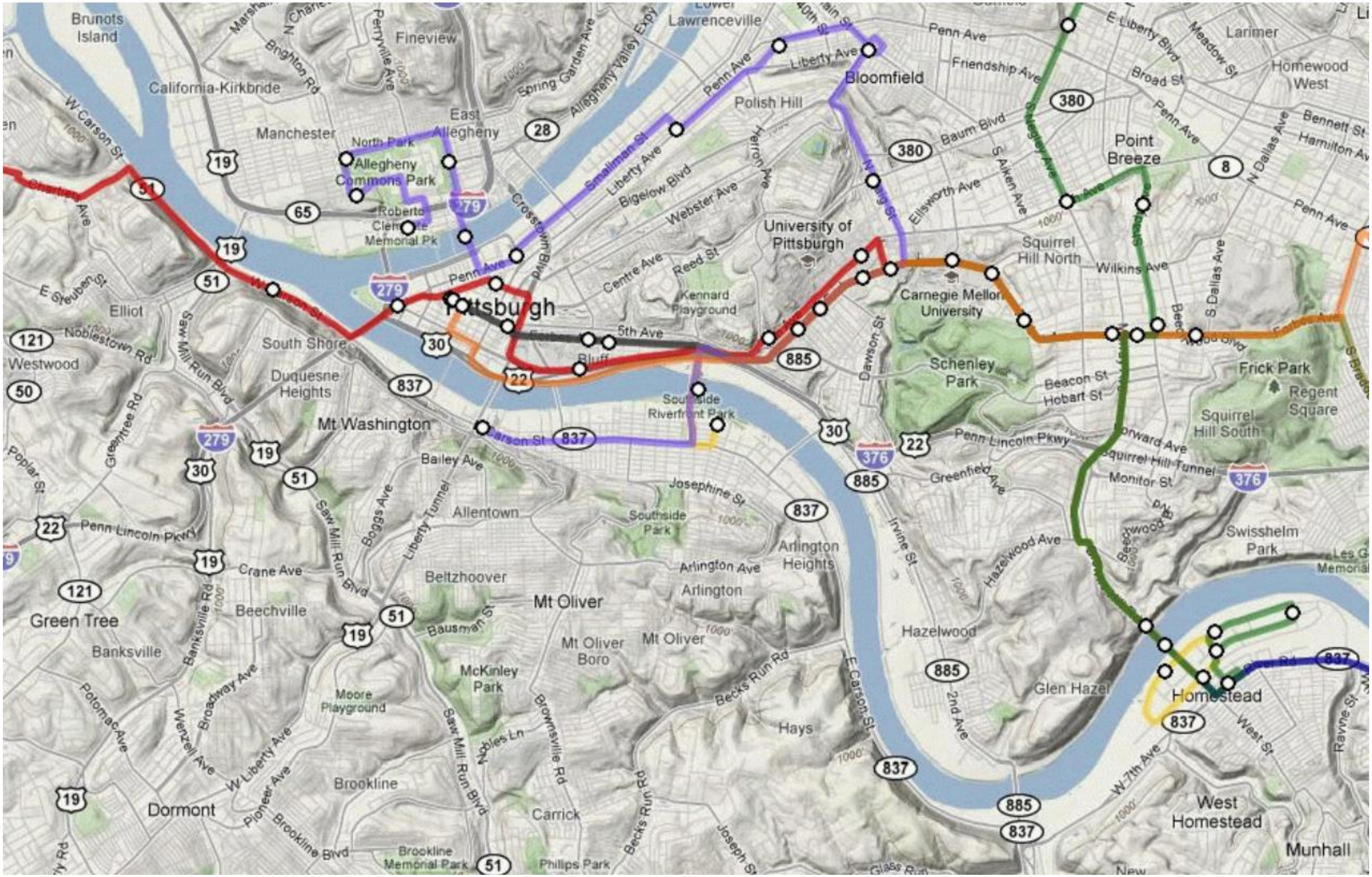
1. Experiments on a fixed corpus may not predict performance in deployment
   - Develop tracker on training data drawn from a particular distribution
   - Deploy tracker into system
   - Tracker causes system to follow a different distribution
   - Problem: train/test mismatch
   - **We explicitly create train/test mismatch in the challenge**
2. Does not directly measure improvement in whole-dialog performance (eg 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

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

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


Brunots Island
Fineview
Lower Lawrenceville
Penn Ave
E Liberty Blvd
Meadow St
Larimer
California-Kirkbride
Brighton Rd
Spring Garden Ave
Allegheny Valley Expy
Liberty Ave
Bloomfield
Friendship Ave
Broad St
Homewood West
Polish Hill
380
Baum Blvd
Penn Ave
Bennett St
Hamilton Ave
Manchester
North Park
East Allegheny
28
Smallman St
Point Breeze
W Carson St
19
Allegheny Commons Park
Liberty Ave
Herron Ave
S Aiken Ave
8
N Dallas Ave
51
65
Roberto Clemente Memorial Pk
279
Bigelow Blvd
Webster Ave
Ellsworth Ave
Penn Ave
Crosstown Blvd
University of Pittsburgh
Squirrel Hill North
E Steuben St
19
51
Penn Ave
Centre Ave
Reed St
Wilkins Ave
S Dallas Ave
279
Elliot
Kennard Playground
Carnegie Mellon University
Pittsburgh
121
Noblestown Rd
Saw Mill Run Blvd
South Shore
30
5th Ave
Bluff
885
Dawson St
Schenley Park
Frick Park
Westwood
Duquesne Heights
Regent Square
50
837
22
Beacon St
Hobart St
Mt Washington
Southside Riverfront Park
Squirrel Hill South
279
19
Carson St
837
30
376
22
Penn Lincoln Pkwy
Forward Ave
Squirrel Hill Tunnel
Poplar St
30
Bailey Ave
51
885
Greenfield Ave
Monitor St
376
Greentree Rd
Josephine St
Boggs Ave
Liberty Tunnel
Allentown
Southside Park
837
Irvine St
Beechwood Blvd
Swisshelm Park
22
Penn Lincoln Pkwy
Hazelwood Ave
19
Crane Ave
Arlington Heights
Arlington Ave
121
Beltzhoover
Saw Mill Run Blvd
Arlington
Green Tree
Beechville
51
Mt Oliver
Banksville
19
Bausman St
Mt Oliver Boro
Mt Oliver
Hazelwood
Banksville Rd
Broadway Ave
McKinley Park
Brownsville Rd
Becks Run Rd
885
Potomac Ave
Moore Playground
E Carson St
Glen Hazel
Homestead
837
Wenzell Ave
W Liberty Ave
Nobles Ln
Hays
2nd Ave
West St
Ravine St
W Liberty Ave
Pioneer Ave
Brookline
Saw Mill Run Blvd
837
W 7th Ave
19
West Homestead
Dormont
Brookline Blvd
Carrick
Becks Run Rd
885
837
Joseph St
Munhall
Brookline Memorial Park
51
Phillips Park
Glass Run

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

Data

Dataset	Source	Calls	Time period	Transcribed?	Labeled?
train1a	Group A	1761	September 2009	Yes	Yes
~~trian1c~~	~~Group A~~	~~1765~~	~~August 2009~~	~~Yes~~	~~Yes~~
train1b *	Group A	14,545	16 Months (2008-2009)	Yes	No
train2	Group A	678	Summer 2010	Yes	Yes
train3	Group B	779	Summer 2010	Yes	Yes
test1	Group A	765	Winter 2011-12	Yes	Yes
test2	Group A	983	Winter 2011-12	Yes	Yes
test3	Group B	1037	Winter 2011-12	Yes	Yes
test4	Group C	451	Summer 2010	Yes	Yes

* 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 includes:
  - 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?	Sorry, which bus route?
Sys dialog acts:	*request(route)*	*sorry(), request(route)*

SLU hyps:

inform(route=61c)	✗
inform(route=28x)	✗
inform(route=61b)	✗

inform(route=56u)	✗
inform(route=61d)	✓

Tracker output (route slot):

inform(route=61c)	0.1	✗
inform(route=28x)	0.3	✗
inform(route=61b)	0.1	✗
none	0.5	✓

inform(route=61c)	0.1	✗
inform(route=28x)	0.1	✗
inform(route=61b)	0.0	✗
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?

schedule	Description
schedule1	Include all turns (regardless of dialog context)
schedule2	Include a turn for a given concept only if : • Concept appears on the SLU N-Best list in that turn, OR • The system's action references that concept in that turn (eg an explicit or implicit confirmation)
schedule3	Include only the last turn of the dialog

Datasets

dataset	Description
`train3.sessions`	All calls in train3
`train3.half1.sessions`	First half of calls in train3
`train3.half2.sessions`	Second half of calls in train3 (encourage participants to report performance by training on half1 and testing on half2)
`train3.call1.sessions`	The first call (for testing)

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
--------------------------------------------------------------------------------
               route from.d from.m from.n to.des to.mon to.nei   date   time  joint
           N    4459   4459   4459   4459   4459   4459   4459   4459   4459   4459
    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

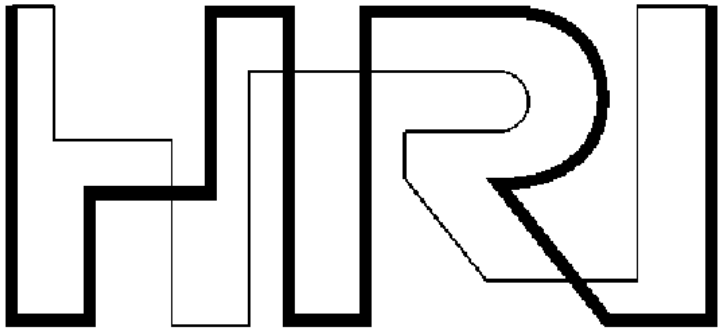
- Test data (in March)

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