Tackling Dialog State Tracking Challenge

Wencan Luo
Department of Computer Science
University of Pittsburgh
PA 15260, USA
wencan@cs.pitt.edu

Abstract

We will propose a model to tackle the Dialog State Tracking Challenge (Williams et al., 2013).

1 Introduction

In dialog systems, "state tracking" refers to accurately estimating the users goal as a dialog progresses. Accurate state tracking is desirable because it provides robustness to errors in speech recognition, and helps reduce ambiguity inherent in language within a temporal process like dialog. Dialog state tracking is an important problem for both traditional uni-modal dialog systems, as well as speech-enabled multi-modal dialog systems on mobile devices, on tablet computers, and in automobiles (Williams et al., 2012).

The 2013 "Dialog State Tracking Challenge"(DSTC) provided a good test bank for this task. This challenge has completed. 9 teams entered a total of 27 entries. Results have been shown at SigDial 2013.

However, the data is still public available[1].

2 Task Description

DSTC data is taken from several different spoken dialog systems. All of them provide bus schedule information for Pittsburgh, Pennsylvania, USA (Black et al., 2013). Different dialog systemw might have different ASR, NLU and dialog control components. In this challenge, only 9 slots are evaluated: route, from.desc, from.neighborhood, from.monument, to.desc, to.neighborhood, to.monument, date, and time. The approximate numbers of distinct values for slots are shown in Table 1. The number of values for each slot varies a lot.

The dialog systems logged SLU N-best hypotheses for each user turn with confidence scores. As they claimed, the coverage of N-best hypotheses is good, so the challenge confines consideration of goals to slots and values that have been observed in an SLU output. The task of a dialog state tracker is to generate a set of observed slot and value pairs, with a score between 0 and 1. The sum of all scores should be 1. In which, the correct slot value should have maximal value.

For evaluation, there are 11 different metrics, 4 test tests under 3 different schedules (Williams et al., 2013) for 9 slots.

3 The Corpus

The data is divided into 4 training sets and 4 test sets. They come from different sources. The basic statistical information for the corpus is shown in Table 2

4 Related Work

4.1 Overall Results

9 teams entered the DSTC, submitting a total of 27 trackers.

Here is the summary of the results (Williams et al., 2013).

Firstly, relative to the baselines, performance on the test data is markedly lower than the training data.

[1] http://research.microsoft.com/en-us/events/dstc/. This link was broken a few days ago, but it is fixed up after a request

Dataset	Source	Calls	Time period	Transcribed?	Labeled?
train1a	Group A	1013	September 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

Table 2: Dataset description

slot name	number of values
route	100
from.desc	500-10000
to.desc	500-10000
from.neighborhood	20-100
to.neighborhood	20-100
from.monument	50-500
to.monument	50-500
date.day	9
date.absmonth	12
date.absday	31
date.relweek	1
time.hour	12
time.minute	60
time.ampm	2
time.arriveleave	2
time.rel	1

Table 1: Approximate number of distinct values for slots

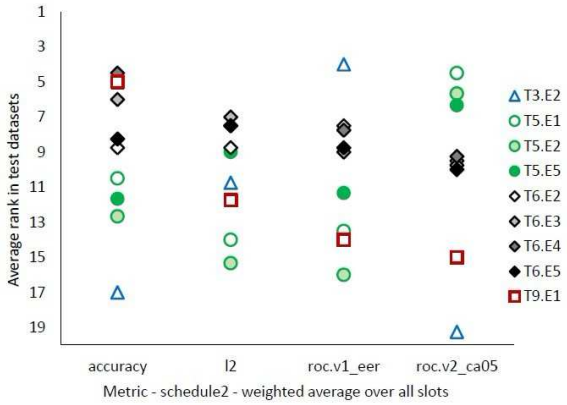

Figure 1: Average rank of top-performing trackers for four metrics. Ranking was done using the given metric, schedule2, and the weighted average of all slots. $Tn.Em$ indicates team n, entry m.

Moreover, only 38% of trackers performed better than a simple majority-class baseline on TEST4.

Secondly, no one wins. Different trackers are tuned for different performance measures, and the optimal tracking algorithm depends crucially on the target performance measure. The average rank of entries for four metrics is shown in Figure 1.

4.2 Methodology

The methods are briefly summarized in Table 3.

Generally speaking, these methods are trying to rescore ASR/NLU engines, but not trying to consider the semantic meaning of the sentences. Commonly features are the rank of NLU slots, the rank of ASR, etc.

Here are some ideas that have not been tried in their reports.

- re-run the NLU
- take the bus schedule into account (the schedule has been given)
- try to really understand the conversation between the computer and the human (Graph Model?)

5 Timeline

Sep 15 - Sep 22

- survey the related work regarding dialog state tracking
- understanding the data, know how to extract and use the data

Team	Author	Method	New Ideas
1	(Metallinou et al., 2013)	MaxEnt	speech recognition error pattern
2	(Henderson et al., 2013)	Deep Neural Network	deep learning
3	(Cuayahuitl et al., 2013)	Bayesian Networks	re-ranking N-best ASR
4	(Lee and Eskenazi, 2013)	MaxEnt	L1 regularization, bins
	(Lee, 2013)	Structured Discriminative model	
5	(Wang and Lemon, 2013)	Rules	infor, deny, affirm, negate
6	(Williams, 2013)	MaxEnt	multi-domain learning
7	(Zilka et al., 2013)	Bayesian, Generative	
8	(Ren et al., 2013)	CRF	
9	(Kim et al., 2013)	Different models	

Table 3: Different approaches submitted by the participants[2]

Sep 23 - Oct 20

- implement discriminative model used the basic features introduced in the SigDial 2013 reports

Oct 21 - Nov 9

- re-run NLU using CRF model
- combined the new NLU with the old NLU

Nov 10 - Dec 12

- do error analysis and propose new features and/or new model for this task

Acknowledgments

Do not number the acknowledgment section.

References

J. D. Williams, A. Raux, D. Ramachandran, and A. W. Black. 2012. *Dialog state tracking challenge handbook*. Technical report, Microsoft Research.

J. D. Williams, A. Raux, D. Ramachandran, and A. W. Black. 2013. *The Dialog State Tracking Challenge*. In Proceedings 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France.

J. Williams. 2013. *Multi-domain learning and generalization in dialog state tracking*. In Proceedings 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France.

A. Black et al. 2011. *Spoken dialog challenge 2010: Comparison of live and control test results*. In Proceedings of SIGDIAL.

H. Cuayahuitl, N. Dethlefs, H. Hastie, O. Lemon. 2013. *Impact of ASR N-Best Information on Bayesian Dialogue Act Recognition*. In Proceedings of SIGDIAL.

M. Henderson, B. Thomson and S. Young. 2013. *Deep Neural Network Approach for the Dialog State Tracking Challenge*. In Proceedings 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France.

D. Kim, J. Choi, K. E. Kim, J. Lee and J. Sohn. 2013. *Engineering Statistical Dialog State Trackers: A Case Study on DSTC*. In Proceedings 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France.

S. Lee and M. Eskenazi. 2013. *Recipe For Building Robust Spoken Dialog State Trackers: Dialog State Tracking Challenge System Description*. In Proceedings 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France.

S. Lee. 2013. *Structured Discriminative Model For Dialog State Tracking*. In Proceedings 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France.

A. Metallinou, D. Bohus, and J. D. Williams 2013. *Discriminative state tracking for spoken dialog systems*. In Proceedings 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France.

H. Ren, W. Xu, Y. Zhang and Y. Yan. 2013. *Dialog State Tracking using Conditional Random Fields*. In Proceedings 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France.

Z. Wang and O. Lemon. 2013. *A Simple and Generic Belief Tracking Mechanism for the Dialog State Tracking Challenge: On the believability of observed information*. In Proceedings 14th Annual Meeting of the Spe-

cial Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France.

L. Zilka, D. Marek, M. Korvas and F. Jurcicek. 2013. *Comparison of Bayesian Discriminative and Generative Models for Dialogue State Tracking*. In Proceedings 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France.