The Dialog State Tracking Challenge Series

Jason D. Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, Deepak Ramachandran

■ *In spoken dialog systems, dialog state tracking refers to the task of correctly inferring the user's goal at a given turn, given all of the dialog history up to that turn. The Dialog State Tracking Challenge is a research community challenge task that has run for three rounds. The challenge has given rise to a host of new methods for dialog state tracking and also to deeper understanding about the problem itself, including methods for evaluation.*

Conversational systems are increasingly becoming a part of daily life, with examples including Apple's Siri, Google Now, Nuance Dragon Go, Xbox, and Cortana from Microsoft, and those from numerous startups. In the core of a conversation system is a key component called a dialog state tracker, which estimates the user's goal given all of the dialog history so far. For example, in a tourist information system, the dialog state might indicate the type of business the user is searching for (pub, restaurant, coffee shop), the desired price range, and the type of food served. 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. At the same time, state tracking is crucial because the system relies on the estimated dialog state to choose actions — for example, which restaurants to suggest. Figure 1 shows an illustration of the dialog state tracking task.

Historically dialog state tracking has been done with hand-crafted rules. More recently, statistical methods have been found to be superior by effectively overcoming some SLU errors, resulting in better dialogs. Despite this progress, direct comparisons between methods have not been possible because past studies use different domains, system components, and evaluation measures, hindering progresss. The Dialog State Tracking Challenge (DSTC) was initiated to address this barrier by providing a common test bed and evaluation framework for dialog state tracking algorithms.

 ISSN 0738-4602

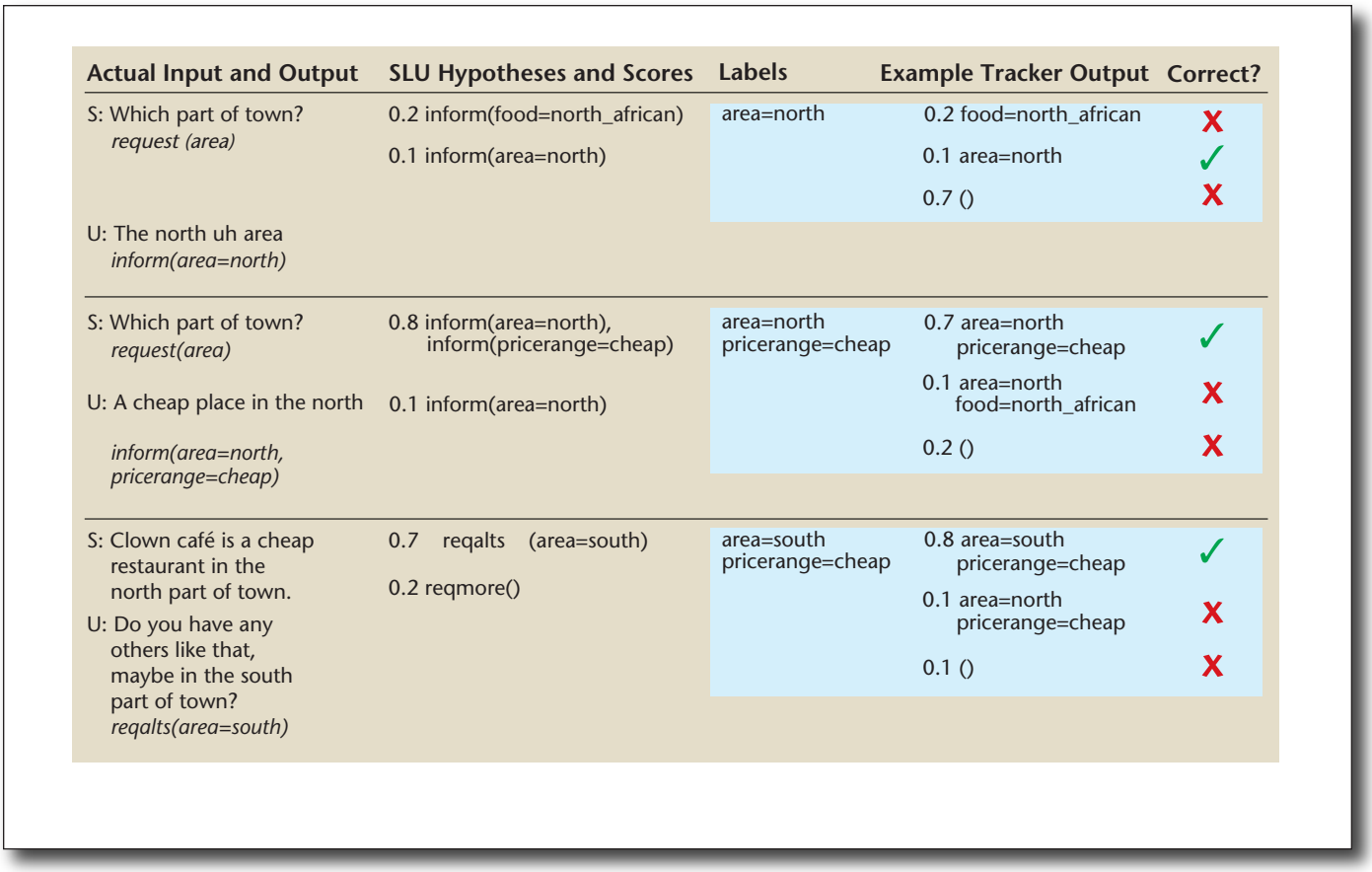

Figure 1. The Dialog State Tracking Problem.

The left column shows the actual dialog system output and user input. The second column shows two SLU n-best hypotheses and their scores. The third column shows the label (correct output) for the user's goal. The fourth column shows example tracker output, and the fifth column indicates correctness.

Challenge Design

The dialog state tracking challenge studies this problem as a corpus-based task. When the challenge starts, labeled human-computer dialogs are released to teams, with scripts for running a baseline system and evaluation. Several months later, a test set of unlabeled dialogs is released. Participants run their trackers, and a week later they return tracker output to the organizers for scoring. After scoring, results and test set labels are made public.

The corpus-based design was chosen because it allows different trackers to be evaluated on the same data, and because a corpus-based task has a much lower barrier to entry for research groups than building an end-to-end dialog system. However when a tracker is deployed, it will inevitably alter the performance of the dialog system it is part of relative to any previously collected dialogs. In order to simulate this mismatch at training time and at run time, and to penalize overfitting to known conditions, dialogs in the test set are conducted using a different dialog manager, not found in the training data.

The first DSTC used 15,000 dialogs between real Pittsburgh bus passengers and a variety of dialog systems, provided by the Dialog Research Center at Carnegie Mellon University (Black et al. 2010). The second and third DSTCs used in total 5,510 dialogs between paid Amazon Mechanical Turkers, who were asked to call a tourist information dialog system and find restaurants that matched particular constraints, provided by the Cambridge University Dialogue Systems Group (Jurcicek, Thomson, and Young 2011).

Each DSTC added new dimensions of study. In the first DSTC, the user's goal was almost always fixed throughout the dialog. In the second DSTC, the user's goal changed in about 40 percent of dialogs. And the third DSTC further tested the ability of trackers to generalize to new domains by including entity types in the test data that were not included in the training data — for example, the training data included only restaurants, but the test data also included bars and coffee shops.

In this relatively new research area, there does not exist a single, generally agreed on evaluation metric; therefore, each DSTC reported a bank of metrics,

sourced from the advisory board and participants. This resulted in approximately 10 different metrics, including accuracy, receiver operating characteristic (ROC) measurements, probability calibration, and so on. Each metric was measured on various subtasks (such as accuracy of a particular component of the user's goal), and at different time resolutions (for example, every dialog turn, just at the end, and so on.) Every combination of these variables was measured and reported, resulting in more than 1000 measurements for each entry. The measurements themselves form a part of the research contribution: after the first DSTC, a correlation analysis was done to determine a small set of roughly orthogonal metrics, which were then reported as featured metrics in DSTC2 and DSTC3, focusing teams' efforts. These featured metrics were accuracy, probability quality (Brier score), and a measure of discrimination computed from an ROC curve.

Each DSTC has been organized by an ad hoc committee, including members of the group providing the dialog data.

Participation and Results

About nine teams have participated in each DSTC, with global representation of the top research centers for spoken dialog systems. Participants have mostly been academic instutions, with a minority of corporate research labs. Results have been presented at special sessions: DSTC1 at the annual Special Interest Group on Discourse and Dialogue (SIGdial) conference in 2013 (Williams et al. 2013); DSTC2 at SIGdial in June 2014 (Henderson, Thomson, and Williams 2014); and DSTC3 at IEEE Spoken Language Technologies (SLT) Workshop in December 2014 (forthcoming).

Papers describing DSTC entries have broken new ground in dialog state tracking; the best-performing entries have been based on conditional random fields (Lee and Eskenazi 2013), recurrent neural networks (Henderson, Thomson, and Young 2014), and web-style ranking (Williams 2014). At present, dialog state trackers are able to reliably exceed the performance of a carefully tuned hand-crafted tracker — for example, in DSTC2, the best trackers achieved approximately 78 percent accuracy versus the baseline's 72 percent. This is impressive considering the maximum performance possible with the provided SLU is 85 percent, due to speech recognition errors.

Prior to the DSTC series, most work on dialog state tracking was based on generative models; however, the most successful DSTC entries have been discriminatively trained models, and these are now the dominant approach. Thus the DSTC series has had a clear impact on the field.

Future Activities

All of the DSTC data will remain available for download, including labels, output from all entries, and the raw tracker output.[1,2] We encourage researchers to use this data for research into dialog state tracking or for other novel uses. In addition, a special issue to the journal *Dialogue and Discourse* will feature work on the DSTC data, and we anticipate publication in 2015. In future challenges, it would be interesting to study aspects of dialog state beyond the user's goal — for example, the user's attitude and expectation. It would also be interesting to consider turn-taking and state tracking of incremental dialogs, where updates are made as each word is recognized. Finally, researchers with dialog data available who would be interested in organizing a future DSTC are encouraged to contact the authors.

Acknowledgements

For DSTC1, we thank the Dialog Research Center at Carnegie Mellon University for providing data, and Microsoft and Honda Research Institute for sponsorship. For DSTC2 and DSTC3, we thank Cambridge University's dialog systems group for providing data. We also thank our advisory committee, including Daniel Boies, Paul Crook, Maxine Eskenazi, Milica Gasic, Dilek Hakkani-Tur, Helen Hastie, Kee-Eung Kim, Ian Lane, Sungjin Lee, Oliver Lemon, Teruhisa Misu, Olivier Pietquin, Joelle Pineau, Brian Strope, David Traum, Steve Young, and Luke Zettlemoyer. Thanks also to Nigel Ward for helpful comments.

Notes

1. research.microsoft.com/events/dstc
2. camdial.org/~mh521/dstc

References

Black, A.; Burger, S.; Langner, B.; Parent, G.; and Eskenazi, M. 2010. Spoken Dialog Challenge 2010. In *Proceedings of the 2010 IEEE Spoken Language Technology Workshop*. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

Henderson, M.; Thomson, B.; and Williams, J. D. 2014. The Second Dialog State Tracking Challenge. In P*roceedings of the 15th Annual SIGdial Meeting on Discourse and Dialogue*. Stroudsburg PA: Association for Computational Linguistics.

Henderson, M.; Thomson, B.; and Young, S. 2014. Word-Based Dialog State Tracking with Recurrent Neural Networks. In *Proceedings of the 15th Annual SIGdial Meeting on Discourse and Dialogue*. Stroudsburg PA: Association for Computational Linguistics.

Jurcicek, F.; Thomson, B.; and Young, S. 2011. Natural Actor and Belief Critic: Reinforcement Algorithm for Learning Parameters of Dialogue Systems Modelled as POMDPs. *ACM Transactions on Speech and Language Processing* 7(3).

Lee, S., and Eskenazi, M. 2013. Recipe for Building Robust Spoken Dialog State Trackers: Dialog State Tracking Challenge System Description. In *Proceedings of the 14th Annual SIGdial Meeting on Discourse and Dialogue*. Stroudsburg PA: Association for Computational Linguistics.

Williams, J. D. 2014. Web-Style Ranking and SLU Combination for Dialog State Tracking. In *Proceedings of the 15th Annual SIGdial Meeting on Discourse and Dialogue*. Stroudsburg PA: Association for Computational Linguistics.

Williams, J. D.; Raux, A.; Ramachadran, D.; and Black, A. 2013. The Dialog State Tracking Challenge. In *Proceedings of the 14th Annual SIGdial Meeting on Discourse and Dialogue*. Stroudsburg PA: Association for Computational Linguistics.

Jason Williams is a researcher at Microsoft Research. His interests include spoken dialog systems, planning under uncertainty, spoken language understanding, and speech recognition. He has published more than 50 journal, conference, and workshop papers and filed more than 10 patents. He is currently vice president of SIGdial — the Special Interest Group on Dialog and Discourse — and served as program cochair for the SIGdial conference in 2013. In the past he has served on the board of directors of the Association for Voice Interaction Design (AVIxD), and on the IEEE Speech and Language Technical Committee (SLTC) in the area of spoken dialog systems. He holds a Ph.D. and master's in speech and language processing from Cambridge University (UK) and a BSE in electrical engineering from Princeton University (USA). Prior to Microsoft, Williams was a principal member of the technical staff at AT&T Labs — Research from 2006–2012.

Matthew Henderson is his Ph.D. under Steve Young at the dialog systems group in Cambridge, UK. His work has looked at statistical methods for dialog systems, particularly in spoken language understanding and dialog state tracking. He studied mathematics at Cambridge for his undergraduate degree, and has a master's degree in speech and language technology from Edinburgh University. He also has a Google Research doctoral fellowship in speech technology.

Antoine Raux is a principal researcher in speech and fusion at Lenovo Inc. His work focuses on various aspects of spoken dialog systems and multimodal human-machine interaction. From 2009 to 2013, Raux was a senior scientist at Honda Research Institute USA, where he led research on human-robot interaction, dialog state tracking, and in-vehicle situated dialog. He is also the main author of the Carnegie Mellon University's Let's Go spoken dialog system, a telephone-based bus schedule information system for the city of Pittsburgh, which has answered more than 200,000 calls from the general public and has been used as a benchmark and a source of real user data throughout the dialog research community. Raux has authored more than 30 papers in peer-reviewed journals and conferences. He holds a Ph.D. in language technologies from Carnegie Mellon University, a master's in intelligence science and technology from Kyoto University (Japan), and an engineering degree from Ecole Polytechnique (France).

Blaise Thomson is chief executive officer and a cofounder of VocalIQ. His interests include robust dialog system design and spoken language understanding and generation. He has published around 40 journal, conference, and workshop papers in the area and completed his Ph.D. and master's degrees in spoken language processing at Cambridge University. Prior to VocalIQ, Thomson was a research fellow at St John's College, Cambridge.

Alan W Black is a professor in the Language Technologies Institute at Carnegie Mellon University (CMU). He did his master's and doctorate at the University of Edinburgh. Before joining the faculty at CMU in 1999, he worked in the Centre for Speech Technology Research at the University of Edinburgh, and before that at ATR in Japan. He is one of the principal authors of the free software Festival Speech Synthesis System and FestVox voice building tools, which constitute the basis for many research and commerical systems around the world. He also works in spoken dialog systems, the Let's Go Bus Information project and mobile speech-to-speech translation systems. Black is an elected member of ISCA board (2007–2015). He has more than 200 refereed publications and is one of the highest cited authors in his field.

Deepak Ramachandran is a research scientist at Nuance Communications Inc. His research interests include dialog management, reinforcement learning, and knowledge representation. He holds a Ph.D. in computer science from the University of Illinois at Urbana-Champaign, and he was previously a scientist at the Honda Research Institute, USA.