A belief tracking challenge task for spoken dialog systems

Jason D. Williams
Microsoft Research, Redmond, WA 98052 USA
jason.williams@microsoft.com

Abstract

Belief tracking is a promising technique for adding robustness to spoken dialog systems, but current research is fractured across different teams, techniques, and domains. This paper amplifies past informal discussions (Raux, 2011) to call for a *belief tracking challenge* task, based on the *Spoken dialog challenge* corpus (Black et al., 2011). Benefits, limitations, evaluation design issues, and next steps are presented.

1 Introduction and background

In dialog systems, *belief tracking* refers to maintaining a distribution over multiple dialog states as a dialog progresses. Belief tracking is desirable because it provides robustness to errors in speech recognition, which can be quite common.

This distribution can be modeled in a variety of ways, including heuristic scores (Higashinaka et al., 2003), Bayesian networks (Paek and Horvitz, 2000; Williams and Young, 2007), and discriminative models (Bohus and Rudnicky, 2006). Techniques have been fielded which scale to realistically sized dialog problems and operate in real time (Young et al., 2009; Thomson and Young, 2010; Williams, 2010; Mehta et al., 2010). In lab settings, belief tracking has been shown to improve overall system performance (Young et al., 2009; Thomson and Young, 2010).

Despite this progress, there are still important unresolved issues. For example, a deployment with real callers (Williams, 2011) found that belief tracking sometimes degraded performance due to model mis-matches that are difficult to anticipate at training time. What is lacking is a careful comparison of methods to determine their relative strengths, in terms of generalization, sample efficiency, speed, etc.

This position paper argues for a belief tracking challenge task. A corpus of labeled dialogs and scoring code would be released. Research teams would enter one or more belief tracking algorithms, which would be evaluated on a held-out test set.

2 Corpus

The *Spoken dialog challenge* corpus is an attractive corpus for this challenge. It consists of phone calls from real (not simulated) bus riders with real (not imagined) information needs. There have been 2 rounds of the challenge (2010, and 2011-2012), with 3 systems in each round. The rounds differed in scope and (probably) user population. A total of 3 different teams entered systems, using different dialog designs, speech recognizers, and audio output. For each system in each round, 500-1500 dialogs were logged. While it would be ideal if the corpus included more complex interactions such as negotiations, as a publicly available corpus it is unparalleled in terms of size, realism, and system diversity.

There are limitations to a challenge based on this corpus: it would not allow comparisons across domains, nor for multi-modal or situated dialog. These aspects could be left for a future challenge. Another possible objection is that off-line experiments would not measure end-to-end impact on a real dialog system; however, we do know that good belief tracking improves dialog performance (Young

NAACL-HLT 2012 Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data, pages 23–24,
Montréal, Canada, June 7, 2012.

et al., 2009; Thomson and Young, 2010; Williams, 2011), so characterizing and improving belief tracking seems a logical next step. Moreover, building an end-to-end dialog system is a daunting task, out of reach of many research teams without specific funding. A corpus-based challenge has a much lower barrier to entry.

3 Evaluation issues

There are many (not one!) metrics to evaluate. It is crucial to design these in advance and implement them as computer programs for use during development. Specific metrics could draw on the following core concepts. **Baseline accuracy** measures the speech recognition 1-best – i.e., accuracy without belief tracking. **1-best accuracy** measures how often the belief tracker's 1-best hypothesis is correct. **Mean reciprocal rank** measures the quality of the ordering of the belief state, ignoring the probabilities used to order; **log-likelihood** measures the quality of the probabilities. **ROC curves** measure the 1-best discrimination of the belief tracker at different false-accept rates, or at the **equal error rate**.

An important question is *at which turns* to assess the accuracy of the belief in a slot. For example, accuracy could be measured at every turn; every turn after a slot is first mentioned; only turns where a slot is mentioned; only turns where a slot appears in the speech recognition result; and so on. Depending on the evaluation metric, it may be necessary to annotate dialogs for the user's goal, which could be done automatically or manually. Another issue is how to automatically determine whether a belief state value is correct at the semantic level.

A final question is how to divide the corpus into a training and test set in a way that measures robustness to the different conditions. Perhaps some of the data from the second round (which has not yet been released) could be held back for evaluation.

4 Next steps

The next step is to form a group of interested researchers to work through the issues above, particularly for the preparation of the corpus and evaluation methodology. Once this is documented and agreed, code to perform the evaluation can be developed, and additional labelling (if needed) can be started.

Acknowledgments

Thanks to Antoine Raux for advocating for this challenge task, and for helpful discussions. Thanks also to Spoken Dialog Challenge organizers Alan Black and Maxine Eskenazi.

References

AW W Black, S Burger, A Conkie, H Hastie, S Keizer, O Lemon, N Merigaud, G Parent, G Schubiner, B Thomson, JD Williams, K Yu, SJ Young, and M Eskenazi. 2011. Spoken dialog challenge 2010: Comparison of live and control test results. In *Proc SIGdial Workshop on Discourse and Dialogue, Portland, Oregon*.

D Bohus and AI Rudnicky. 2006. A 'K hypotheses + other' belief updating model. In *Proc AAAI Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems, Boston*.

H Higashinaka, M Nakano, and K Aikawa. 2003. Corpus-based discourse understanding in spoken dialogue systems. In *Proc ACL, Sapporo*.

N Mehta, R Gupta, A Raux, D Ramachandran, and S Krawczyk. 2010. Probabilistic ontology trees for belief tracking in dialog systems. In *Proc SIGdial Workshop on Discourse and Dialogue, Tokyo, Japan*.

T Paek and E Horvitz. 2000. Conversation as action under uncertainty. In *Proc Conf on Uncertainty in Artificial Intelligence (UAI), Stanford, California*, pages 455–464.

A Raux. 2011. Informal meeting on a belief tracking challenge at interspeech.

B Thomson and SJ Young. 2010. Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems. *Computer Speech and Language*, 24(4):562–588.

JD Williams and SJ Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):393–422.

JD Williams. 2010. Incremental Partition Recombiantion for Efficient Tracking of Multiple Dialogue States. In *ICASSP*, Dallas, TX.

JD Williams. 2011. An empirical evaluation of a statistical dialog system in public use. In *Proc SIGDIAL, Portland, Oregon, USA*.

SJ Young, M Gašić, S Keizer, F Mairesse, J Schatzmann, B Thomson, and K Yu. 2009. The hidden information state model: a practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language*.