Comparative Error Analysis of Dialog State Tracking

Ronnie W. Smith
Department of Computer Science
East Carolina University
Greenville, North Carolina, 27834
rws@cs.ecu.edu

Abstract

A primary motivation of the Dialog State Tracking Challenge (DSTC) is to allow for direct comparisons between alternative approaches to dialog state tracking. While results from DSTC 1 mention performance limitations, an examination of the errors made by dialog state trackers was not discussed in depth. For the new challenge, DSTC 2, this paper describes several techniques for examining the errors made by the dialog state trackers in order to refine our understanding of the limitations of various approaches to the tracking process. The results indicate that no one approach is universally superior, and that different approaches yield different error type distributions. Furthermore, the results show that a pairwise comparative analysis of tracker performance is a useful tool for identifying dialogs where differential behavior is observed. These dialogs can provide a data source for a more careful analysis of the source of errors.

1 Introduction

The Dialog State Tracking Challenge (Henderson et al., 2013) provides a framework for comparison of different approaches to tracking dialog state within the context of an information-seeking dialog, specifically information about restaurants in the Cambridge, UK, area. The challenge makes available an annotated corpus that includes system logs from actual human-machine dialog interactions. These logs include information about the system dialog acts, the N-best speech recognition hypotheses, and the hypothesized interpretation (including confidence estimates) of the user's spoken utterances as provided by the dialog system's Spoken Language Understanding (SLU) module.

Consequently, standalone algorithms for tracking the state of the dialog can be developed and tested. While performance as part of an actual dialog interaction cannot easily be evaluated (because differing results produced by different trackers may lead to different choices for system dialog acts in a real-time interaction), performance on a turn-by-turn basis can be evaluated and compared.

Results from the first challenge were presented in several papers at SIGDial 2013 (general reference Williams et al. (2013)) and highlighted several different approaches. These papers focused on comparative performance as well as a description of the various techniques for tracking dialog state that were employed. However, there was no detailed error analysis about tracker performance, either within or across trackers. Such analysis can help further our understanding of the sources and impact of dialog miscommunication. This paper presents such an analysis from the current Dialog State Tracking Challenge (DSTC 2) using the publicly available results of the challenge (http://camdial.org/~mh521/dstc/). This paper describes techniques for examining the following aspects of performance as it relates to tracking errors and their potential impact on effective communication in dialog.

- Estimating an upper bound on accuracy.
- Error distribution as a function of tracker—both globally and subdivided by acoustic model or attribute type.
- Pairwise comparative accuracy of trackers—for what types of dialogs does one tracker perform better than another?

Initial results based on application of these techniques are also presented.

Proceedings of the SIGDIAL 2014 Conference, pages 300–309,
Philadelphia, U.S.A., 18-20 June 2014.

2 Data Source: DSTC 2

DSTC 2 is based on corpora collected on dialogs about restaurant information for Cambridge, UK. Besides introducing a different domain from the original DSTC (that dealt with bus timetables) DSTC 2 is structured in such a way as to allow for the possibility of changing user goals and thus represents a more significant challenge for dialog state tracking. An overview of the current challenge and results can be found in Henderson et al. (2014).

2.1 Nature of Dialogs

Unlike the dialogs of the original DSTC that were based on actual uses of the bus timetable information system, the dialogs for DSTC 2 were collected in the more traditional experimental paradigm where system users were given a dialog scenario to follow. Example scenario descriptions extracted from two of the log files are given below.

- `Task 09825: You want to find a cheap restaurant and it should be in the south part of town. Make sure you get the address and phone number.`

- `Task 05454: You want to find an expensive restaurant and it should serve malaysian food. If there is no such venue how about korean type of food. Make sure you get the address and area of the venue.`

The basic structure of the dialogs has the following pattern.

1. Acquire from the user a set of constraints about the type of restaurant desired. Users may supply constraint information about area, food, name, and price range. This phase may require multiple iterations as user goals change.

2. Once the constraints have been acquired, provide information about one or more restaurants that satisfy the constraints. Users may request that additional attributes about a restaurant be provided (such as address and phone number).

2.2 Measuring Task Performance

Because of the complex nature of statistical dialog state tracking there are many different reasonable ways to measure tracker performance. Besides evaluating the accuracy of the 1-best hypothesis there are also a number of possible measures based on the quality of the estimate for dialog state (see Henderson et al. (2013) for details).

For the purpose of this paper the analysis will be based on tracker performance on accuracy (1-best quality) for the joint goal based on the four previously mentioned constraint attributes (area, food, name, and price range). The reason for this choice is that in an actual human-system dialog in an information-seeking domain, the dialog manager must choose an action based on the system's beliefs about the constraining attributes. While level of belief might positively influence when to engage in explicit or implicit confirmation, ultimate success depends on correct identification of values for the constraining attributes. Having too much confidence in inaccurate information has always been a major error source in dialog systems. Consequently, 1-best joint goal accuracy is the focus of study in this paper.

2.3 Description of Error Types

Since we are focused on joint goal accuracy, error type classification will be based on the following three types of possible deviations from the true joint goal label for a given turn.

1. Missing Attributes (MA) - these are situations where a value for an attribute has been specified in the actual data (e.g. "belgian" for the attribute "food"), but the dialog state tracker has no value for the attribute in the joint belief state.[1]

2. Extraneous Attributes (EA) - these are situations where the tracker has a value for the attribute in the joint belief state, but the attribute has not been mentioned in the actual dialog.

[1] The format of DSTC 2 allows for automatic compilation of the joint belief state by the scoring software. The probability mass for a given attribute that is not assigned to specific values for attributes is assigned to a special value *None*. If no value for the attribute has a probability estimate exceeding *None*, then no value for that attribute is included in the joint belief state. It is also possible for a dialog state tracker to explicitly provide a joint belief state. In DSTC 2 some systems do explicitly provide a joint belief state while others use the default.

3. False Attributes (FA) - these are situations where a value for an attribute has been specified in the actual data (e.g. "catalan" for the attribute "food"), but the dialog state tracker has a different value (such as "fusion" for "food").

For turns where there are errors, it is certainly possible that multiple errors occur, both multiple errors of a given type, and multiple errors of different types. This is taken into consideration as described next.

2.4 Recording Tracker Performance

For each tracker a data file consists of a sequence of tuples of the form ($Correct$,EA,MA,FA) that were generated for each turn for which there was a valid joint goal label.[2]. The meaning of each value in the tuple is given below.

- $Correct$ - has the value 1 if the tracker joint goal label is correct and 0 if it was incorrect.

- EA - a count of the number of different extraneous attributes that occurred in the turn. Will always be 0 if $Correct$ = 1.

- MA - a count of the number of different missing attributes that occurred in the turn. Will always be 0 if $Correct$ = 1.

- FA - a count of the number of different false attributes that occurred in the turn. Will always be 0 if $Correct$ = 1.

Consequently, whenever $Correct$ is 1, the tuple will always be of the form (1,0,0,0). If $Correct$ is 0, at least one of the three following entries in the tuple will have a non-zero value.

These files were generated by modifying the scoring script provided by the DSTC organizing committee. The modification causes the necessary information to be output for each relevant turn. These data files represent the result of tracker performance on 1117 dialogs over a total of 9689 turns.

2.5 Mapping Labels to Dialogs

Another modified version of the scoring script was used to iterate through the dialogs to produce a template that associates each of the 9689 labeled turns with the specific (dialog ID, turn within dialog) pair that the turn represents. This information was used in the error analysis process to identify specific dialogs for which tracking was not particularly accurate (see section 4).

2.6 Choice of Trackers

There were a total of 9 different teams that submitted a total of 31 trackers for DSTC 2. For this study, one tracker from each team is being used. The choice of tracker is the one that performed the best on 1-best joint goal accuracy, one of the overall "featured metrics" of the challenge (Henderson et al., 2013). Their performance on this metric ranged from 50.0% to 78.4%. Seven of the nine trackers had performance of better than 69%, while there were two performance outliers at 50% and 60%.

For purposes of this study, it seemed best to include a tracker from all groups since part of the intent of the challenge is to carefully examine the impact of different approaches to dialog state tracking. Based on the optional descriptions that teams submitted to the challenge, there were quite a variety of approaches taken (though not all teams provided a description). Some systems used the original SLU results. Other systems ignored the original SLU results and focused on the ASR hypotheses. Some systems created their own modified versions of the original SLU results. Modeling approaches included Maximum Entropy Markov model, Deep Neural Network model, rule-based models, Hidden Information State models, and conditional random fields. Hybrid approaches were used as well. A few more details about our submitted tracker will be provided in section 4.

One of the purposes of this study was to look at the distribution of errors based on the different types discussed in section 2.3, both in absolute and relative terms. Consequently, one intended investigation is to see if there is a difference in error type distribution depending on a number of parameters, including the approach used to dialog state tracking. Thus, examining the results from the top trackers of all teams can provide valuable information regardless of the absolute accuracy of the

[2] In some cases at the start of dialogs, no SLU hypotheses have yet to mention any values for any of the joint goal attributes. As mentioned in Henderson et al. (2013), those turns are not included in the joint-goal accuracy evaluation. This occurred in a total of 201 turns over 193 dialogs.

tracker. As it turned out, each tracker studied had multiple turns where it was the only tracker to provide a correct joint goal label. This happened on about 4% of all the turns. The number of turns for which a tracker was the only tracker to provide a correct joint goal label ranged from 5 to 89 and tended to follow the general ranking of accuracy (i.e., more accurate trackers tended to have more turns where it was the only tracker correct). However, it did not follow the relative rankings precisely.

3 Analysis: Global Tracker Performance

3.1 How much misunderstanding can be expected?

Another way to ask this question would be, "what error rate should be expected from a high performance tracker? For example, there were 21 dialogs consisting of 8 user turns or more where none of the trackers under study correctly represented the joint goal for any turn.

Looking more broadly, there were 1332 turns over the entire set of dialogs for which none of the trackers had a correct representation of the joint goal. Thus, if we could construct an "oracle" tracker that could always select the correct representation of the joint goal from among the nine trackers under study (when at least one of them had the correct representation), this would imply an error rate of 13.7%.[3] This contrasts with an error rate of 21.6% for the best performing tracker submitted as part of DSTC 2. If we look at tracker performance as a function of acoustic model (artificially degraded (A0), and optimized (A1)), the error rate estimate for the oracle tracker is 17.0% using model A0 and 10.3% using model A1.

3.2 Global Error Type Distribution

Using the classification of error types described in section 2.3: Extraneous Attributes (EA), Missing Attributes (MA), and False Attributes (FA), we can explore the distribution of error types as a function of the dialog tracker. Table 1 provides a summary view of the distributions over all the dialogs of the test set. For comparison, the baseline focus tracker provided by the DSTC 2 organizers (see Henderson et al. (2013)) and the HWU baseline tracker provided by Zhuoran Wang of Heriot-Watt University (see `http://camdial.org/~mh521/dstc/`) are also included. While trackers 1 and 9 are also presented for completeness, the main focus of the analysis is on trackers 2 through 8, the trackers with higher levels of performance on the featured metric of 1-best joint goal accuracy. Each row represents the relative distribution of errors by a given tracker. For example, for our tracker, tracker 3, there were 2629 turns (out of the total 9689 turns) where the tracker made one or more errors for the attributes of the joint goal. There were a total of 3075 different attribute errors of which 545 or 17.7% of the errors were of type EA, 1341 or 43.6% were of type MA, and 1189 or 38.7% of type FA. A visual representation of this information is provided in the Appendix in figure 1. Some general observations are the following.

- Other than tracker 5, the relative number of errors of type MA exceeded the relative number of errors of type FA. For attributes actually mentioned by the user, trackers in general were more likely to reject a correct hypothesis (leading to a type MA error) than accept an incorrect hypothesis (leading to a type FA error).

- Based on the brief description provided with submission of the tracker, tracker 5 uses a hybrid approach for tracking the different goals (one of the baseline trackers for the food attribute, but an n-best approach to the others). This approach seemed to lead to the acceptance of more spurious hypotheses than the other trackers (hence the higher EA rate). Tracker 8 also had a slightly higher error rate for EA. Its submission description indicates the combined use of several models, at least one of which used the training data for developing model parameters.

3.3 Error Type Distribution as a Function of Acoustic Model

Since publicly available spoken dialog systems cannot control the environment in which they are used, speech recognition rates can vary widely. One of the general goals of the DSTC is to evaluate tracker performance for varying levels of speech recognition accuracy. Hence the use in

[3] Note that this is not any sort of an absolute estimate. For example, if provided baseline trackers are included (one provided by the DSTC 2 organizers and another by Zhuoran Wang of Heriot-Watt University), the number of turns where no tracker correctly represents the joint goal reduces to 1325 turns.

	Total Errors		EA		MA		FA	
Tracker	# Turns	# Errors	Count	Percent	Count	Percent	Count	Percent
Focus	2720	3214	652	20.3%	1124	35.0%	1438	44.7%
HWU	2802	3352	601	17.9%	1526	45.5%	1225	36.6%
1	3865	4411	673	15.3%	2436	55.2%	1302	29.6%
2	2090	2432	451	18.5%	1177	48.4%	804	33.1%
3	2629	3075	545	17.7%	1341	43.6%	1189	38.7%
4	2246	2598	441	17.0%	1100	42.3%	1057	40.7%
5	2956	3618	947	26.2%	1218	33.7%	1453	40.2%
6	2730	3231	552	17.1%	1410	43.6%	1269	39.3%
7	2419	2791	446	16.0%	1205	43.2%	1140	40.8%
8	2920	3546	763	21.5%	1456	41.0%	1327	37.4%
9	4857	6183	781	12.6%	4222	68.3%	1180	19.1%

Table 1: Error Distribution: all dialogs

DSTC 2 of two acoustic models: model A1 which is a model optimized for the domain, and model A0 which has artificially degraded acoustic models (Henderson et al., 2013). For the test set, there were 542 dialogs yielding 4994 turns with joint goal labels for model A0, and 575 dialogs yielding 4695 turns with joint goal labels for model A1. It is unsurprising that the average number of turns in a dialog was shorter for the dialogs using the more accurate speech recognizer.

The previous table looked at the global behavior combining all the dialogs. An interesting question to examine is if the error distributions change as a function of acoustic model. Tables 2 and 3 give some insight into that question. Table 2, the results using the optimized model A1, unsurprisingly shows that when the speech signal is better and by implication the SLU confidence scores are stronger and more accurate, the relative rate of type FA errors declines while the relative rate of type MA errors increases (when compared to the overall results of Table 1). For errors of type EA it is about an even split—for some the relative number of EA errors decreases, and for some it increases. The results in Table 3 for the A0 model show the opposite trend for the relative errors of type MA compared to type FA.

3.4 Error Type Distribution as a Function of Attribute

While it is future work to do an exact count to determine the frequency with which the four different constraining attributes (area, food, name, and price range) are actually mentioned in the dialogs, it is clear from the data that the primary objects of conversation are area, food, and price range. This makes sense, since there are often alternative effective ways to access information about a restaurant other than to interact with a dialog system given that the name has already been determined by the user.[4] Consequently, for the remaining three attributes, an investigation into the relative distribution of errors as a function of attribute type within error type was conducted. The results are presented in Table 4. This table is looking at all the test data combined and not separating by acoustic model. Again the focus of discussion will be trackers 2 through 8. For brevity, the results for error type FA are omitted as they are pretty similar for all trackers.

> relative error rate for food >> than relative error rate for area >> than relative error rate for price range.

This follows naturally from the fact that there are 91 possible values for food, 5 possible values for area, and only 3 possible values for price range. Thus, there are many more possibilities for confusion for the value for the food attribute. When we examine the results in Table 4, there are a variety of interesting observations.

- Within error type EA, the only trackers for which the relative error rate for price range exceeds the relative error rate for area are trackers 5 and 7.
- Trackers 3 and 4 are more prone to have EA errors for the food attribute.

[4] One of the anonymous reviewers pointed out that the choice of scenarios used in the data collection process is also a factor.

Tracker	EA	MA	FA
1	12.4%	61.7%	25.9%
2	20.5%	53.4%	26.1%
3	16.1%	50.3%	33.7%
4	17.8%	45.7%	36.6%
5	25.7%	40.9%	33.4%
6	17.5%	49.9%	32.6%
7	15.0%	53.6%	31.5%
8	23.0%	43.0%	34.0%
9	11.6%	71.7%	16.7%

Table 2: Error Distribution: A1 dialogs

Tracker	EA	MA	FA
1	17.3%	50.5%	32.1%
2	17.5%	45.8%	36.6%
3	18.7%	39.8%	41.5%
4	16.5%	40.4%	43.0%
5	26.4%	29.9%	43.7%
6	16.8%	40.0%	43.2%
7	16.5%	37.4%	46.0%
8	20.6%	39.9%	39.5%
9	13.3%	65.8%	20.8%

Table 3: Error Distribution: A0 dialogs

- Trackers 2, 6, 7, and 8 all have a noticeable jump in the relative error rate for the food attribute for type MA errors over type EA errors. In contrast, trackers 3, 4, and 5 show a noticeable decrease.

What of course is missing from these observations is any conjecture of causality based on a careful analysis of individual tracker behavior. Given the lack of accessibility to the details of system implementations for all the trackers, other techniques of investigation are needed. The next section explores another potentially valuable technique—comparing the results of two trackers on a turn-by-turn basis, and using these results to identify particular dialogs that exhibit radically different outcomes in performance.

4 Analysis: Pairwise Comparative Accuracy

Another avenue of analysis is to directly compare the performance of two trackers. How do they differ in terms of the types of dialog situations that they handle effectively? We will examine these issues through comparison of the top performing tracker in the challenge (with respect to the featured metric 1-best joint goal accuracy) with our tracker entry, Pirate.[5]

4.1 Pirate methodology: what should dialog expectation mean?

The overarching philosophy behind the development of Pirate is simply the following.

> There is belief about what we think we know, but there should also be an expectation about what comes next if we are correct.

One of the first dialog systems to make use of a hierarchy of dialog expectations was the Circuit Fix-It Shop (Smith et al., 1995) which was also one of the first working dialog systems to be carefully and extensively evaluated (Smith and Gordon, 1997) and (Smith, 1998). However, at the time, the ability to make use of large corpora in system development was largely non-existent.[6]

Our approach in DSTC 2 for making use of the extensive training data combined the SLU hypotheses with confidence scores (interpreted as probabilities) with a simple statistical model of dialog expectation to create modified SLU confidence scores. The model of dialog expectation was based on a simple bigram model using frequency counts for (system dialog act, user speech act) pairs. This can be normalized into a probabilistic model that gives the probability of a user speech act given the context of the most recent system dialog act to which the user is responding. The equation used to modify SLU confidence scores is the following. Let $Prob(SLU)$ represent the confidence score (expressed as a probability) for the hypothesis SLU, and let $Val(SLU)$ represent the actual hypothesis (e.g. inform(food = belgian)).

$$Prob(SLUmod) = 0.7{*}Prob(SLU){+}0.3{*}Expct$$

where $Prob(SLU)$ is the original confidence score for the hypothesis, and $Expct$ is the probability of the occurrence of the speech act used

[5] The mascot name of East Carolina sports teams is the Pirates. In addition, the code development process for our tracker was based on modification of the simple baseline tracker provided by the DSTC 2 organizers.

[6] Moody (1988) used the Wizard-of-Oz paradigm to collect dialogs relevant to the Circuit Fix-It Shop domain as part of her research into the effects of restricted vocabulary on discourse structure, but the total number of dialogs was about 100. In contrast, DSTC 2 provided 1612 actual human-computer dialogs for the training set, 506 dialogs for the development set, and 1117 dialogs for the test set.

	EA			MA		
Tracker	Food	Price	Area	Food	Price	Area
1	41.2%	25.6%	29.9%	41.5%	36.7%	20.2%
2	29.5%	34.8%	35.7%	41.4%	26.6%	27.8%
3	36.0%	23.7%	33.0%	30.9%	34.1%	31.9%
4	44.4%	25.8%	27.7%	30.4%	35.4%	29.8%
5	32.2%	40.5%	27.0%	19.5%	34.0%	42.4%
6	28.8%	34.2%	37.0%	46.0%	27.7%	22.8%
7	26.7%	38.3%	31.4%	35.7%	33.1%	28.0%
8	28.3%	25.7%	44.2%	49.4%	28.9%	18.9%
9	28.3%	17.7%	34.3%	54.5%	17.8%	26.8%

Table 4: Error Distribution by Attribute

in the SLU hypothesis given the current system speech act (i.e., the probability that comes from the statistical model of dialog expectation). The 0.3 weighting factor was determined through trial and error to perform the best given the training data (basing performance on 1-best joint goal accuracy).[7]

After calculating the modified values, the scores are renormalized so that the confidence values sum to 1. Given the renormalized values for $Prob(SLUmod)$, dialog state was updated by using the following rules. Let $Val(HypCur)$ represent the current hypothesis in the dialog state for the value of an attribute, and its confidence score be denoted by $Prob(HypCur)$.

1. Increase $Prob(SLUmod)$ by $Prob(X)$ where $Val(X) == NULL$ (i.e. the default NULL hypothesis for the SLU), whenever $Prob(X)$ is < 0.5. Reset $Prob(X)$ to 0.

2. Replace $HypCur$ with the highest scoring $SLUmod$ for that attribute if the user speech act is an *inform*, and the following relationship holds.

   $$Prob(SLUmod) + Tol \geq Prob(HypCur)$$

   where Tol is an experimentally determined tolerance value (currently set at 0.1).

3. If the system speech act was a *canthelp* act that specifies particular attribute values (e.g. food = belgian), and the current chosen hypothesis ($SLUmod$) provides information about that attribute, overwrite the state information for the attribute listed in *canthelp* even if the confidence score is less.

The motivation for these rules comes from the assumption that the Gricean Cooperative Principle for conversation (Grice, 1975) applies to this dialog environment. Given this assumption, rule 1 is based on the belief that the human dialog participant is attempting to make an appropriate dialog contribution at every turn. Consequently when reasonable, we will augment the top hypothesized SLU's confidence score with any weight given to the NULL hypothesis. Rule 2 is based on the idea that an intended new contribution should replace a previous contribution and that some allowance should be made for "signal noise" in calculating SLU confidence. Rule 3 reflects the idea that when the system cannot provide assistance about a specified attribute value, any new information about the attribute should be considered a replacement.

The above rules are for updating choices for the individual attributes that are possible components of the goal state (area, food, name and price range). In our modeling of dialog state, we only maintain the top actual hypothesis for each attribute, For producing the joint goal, we used the default that the joint goal is the product of the marginal goals.[8]

With this fairly simple approach, Pirate had a 1-best joint goal accuracy of 72.9%. This accuracy rate exceeded the performance of all baseline trackers, and was 13th out of 31 for the trackers submitted.[9]

[7] For recent work using a Bayesian-based alternative for combining dialog history with the current utterance to calculate probabilities, see Raux and Ma (2011).

[8] Consequently, if our confidence score for the top hypothesis is < 0.5, that hypothesis will not be included in the joint goal, as the default "None" is associated with higher confidence.

[9] The set of 12 trackers that performed better is comprised of 4 trackers each from 3 other teams.

4.2 Comparison to the Best Performing Tracker

An entry from team2 achieved 78.4% accuracy on the 1-best joint goal accuracy metric. A comparative analysis was conducted whereby the performance of each tracker was compared on a turn-by-turn basis. Highlights of this analysis include the following.

- The two trackers were both correct 70.6% of the time and both incorrect 19.3% of the time.
- 7.8% of the time Pirate was incorrect when the team2 tracker was correct.
- 2.2% of the time, Pirate was correct when the team2 tracker was incorrect.

Further exploration examined performance within dialogs. It was discovered that there were 18 dialogs where Pirate was incorrect for at least 8 turns where the team2 tracker was correct. Furthermore, there were no turns in those dialogs where the team2 tracker was incorrect when Pirate was correct. Given that the team2 tracker performed several percentage points better overall, this is not surprising. What might be surprising is that there are 7 dialogs where the opposite was true, and Pirate performed better than the team2 tracker. An initial glance at an actual dialog from each situation indicated the following.

- While team2 did not offer a description of their methodology in their submission, it can be inferred that they used the original ASR hypotheses as part of its dialog state tracking. Pirate was unable to detect in the 2nd turn that the goal (area=dontcare) was being communicated because it did not show up in the SLU hypotheses. However, the top ASR hypothesis was "area". Integrating SLU with dialog context is known to be a good idea when technically feasible, and is borne out by this example. This missing attribute for goal state was propagated throughout all subsequent turns of the dialog. However, it should be noted that omitting an attribute where the correct value is "dontcare" is a somewhat benign error as discussed in the next example.
- The dialog reviewed where the team2 tracker had trouble that Pirate did not revolved around the fact that at an important moment in the dialog, the team2 tracker added an unstated hypothesis of the form (food=dontcare) to its joint goal. This was retained for the duration of the dialog. It can be readily argued that this is a benign error. If the user never explicitly gave a constraint about food (implying that $None$ is the correct value for the attribute), the dialog manager is not likely to make a wrong decision if it's basing its action instead on (food=dontcare).

Time constraints have prohibited further examination of the other dialogs, but clearly this is a fruitful area of exploration for understanding behavioral differences between approaches to dialog state tracking.

5 Conclusion

A primary motivation of the DSTC is to allow for direct comparisons between alternative approaches to dialog state tracking. The results from DSTC 1 focused on performance aspects without providing a detailed analysis of errors sources. This paper describes several techniques for examining the errors made by the dialog state trackers in order to refine our understanding of the limitations of various approaches to the tracking process.

Though the analysis at this point is incomplete, one immediate observation is that no one approach is universally superior to other approaches with respect to the performance metric 1-best joint goal accuracy. However, being able to carefully determine the conditions under which one approach outperforms another and determining if there are ways to combine alternative techniques into a more effective but sufficiently efficient tracking model is very much an unsolved problem. The results from this paper suggest that a careful analysis of errors can provide further insight into our knowledge about the difficult challenge of dialog state tracking. We would like to explore some of the trends observed with appropriate statistical tests as well as look more carefully at dialogs where pairwise comparative analysis indicates highly differential behavior.

Acknowledgments

The author would like to acknowledge the valuable work of East Carolina graduate student Ryan Dellana in developing several tools that assisted with compilation of the statistical based expectations as well as tools for generating selected transcripts from the DSTC 2 corpus. A special thanks goes to the organizers of the DSTC 2 challenge for making this work possible. Thanks also go to the anonymous reviewers for their constructive comments. Their suggestions have been very helpful in producing the final version of this paper.

References

H. P. Grice. 1975. Logic and conversation. In P. Cole and J.L. Morgan, editors, *Syntax and Semantics, Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

M. Henderson, B. Thomson, and J. Williams, 2013. *Dialog State Tracking Challenge 2 & 3*. http://camdial.org/~mh521/dstc/downloads/handbook.pdf, accessed March 5, 2014.

M. Henderson, B. Thomson, and J. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the SIGdial 2014 Conference*, Philadelphia, U.S.A., June.

T. S. Moody. 1988. *The Effects of Restricted Vocabulary Size on Voice Interactive Discourse Structure*. Ph.D. thesis, North Carolina State University.

A. Raux and Y. Ma. 2011. Efficient probabilistic tracking of user goal and dialog history for spoken dialog systems. In *Proceedings of INTERSPEECH-2011*, pages 801–804.

R.W. Smith and S.A. Gordon. 1997. Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialog. *Computational Linguistics*, 23:141–168.

R.W. Smith, D.R. Hipp, and A.W. Biermann. 1995. An architecture for voice dialog systems based on Prolog-style theorem-proving. *Computational Linguistics*, 21:281–320.

R.W. Smith. 1998. An evaluation of strategies for selectively verifying utterance meanings in spoken natural language dialog. *International Journal of Human-Computer Studies*, 48:627–647.

J. Williams, A. Raux, D. Ramachandran, and A. Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France, August. Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-4065.

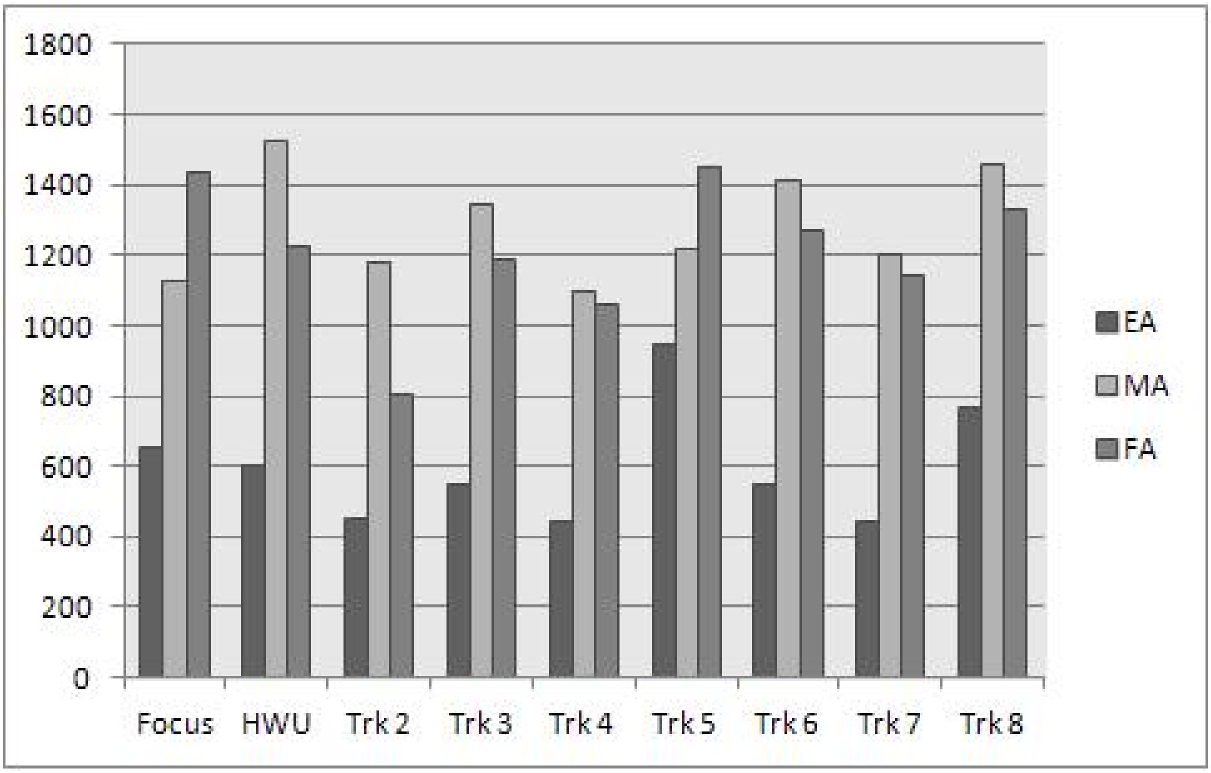

Figure 1: Error Distribution: all dialogs

Appendix

Figure 1 displays in a graphical fashion the error counts for the different types of missing attributes for the trackers listed in Table 1. For clarity, the data for trackers 1 and 9 are omitted. “Focus” is the baseline focus tracker provided by the DSTC 2 organizers (Henderson et al., 2013), and “HWU” is the baseline tracker provided by Zhuoran Wang (see `http://camdial.org/~mh521/dstc/`). “Trk 3” is our tracker, Pirate. As a reminder, the best overall performing tracker is the one labeled “Trk 2”. One observation from the figure is that its best performance is in minimizing False Attribute (FA) errors.