A critical analysis of two statistical spoken dialog systems in public use

Jason D. Williams

Microsoft Research, Redmond, WA, USA
jason.williams@microsoft.com

ABSTRACT

This paper examines two statistical spoken dialog systems deployed to the public, extending an earlier study on one system [1]. Results across the two systems show that statistical techniques improved performance in some cases, but *degraded* performance in others. Investigating degradations, we find the three main causes are (non-obviously) inaccurate parameter estimates, poor confidence scores, and correlations in speech recognition errors. We also find evidence for fundamental weaknesses in the formulation of the model as a generative process, and briefly show the potential of a discriminatively-trained alternative.

1. INTRODUCTION

For more than a decade, researchers have worked to apply statistical techniques to spoken dialog systems. One of the main aims is to improve robustness to errors in automatic speech recognition by maintaining a distribution over many possible hypotheses for the true state of the dialog [2]. In 2010, these techniques were deployed to the general public for the first time, in the first Spoken Dialog Challenge [3]. An analysis found statistical techniques only sometimes improved accuracy, and suggested several improvements [1]. In 2011, these changes were made and re-deployed to the public in a second round of the Spoken Dialog Challenge.

This paper provides a critical analysis spanning these two deployments. The contribution is not a new technique or algorithm, but rather a thorough evaluation of state-of-the-art technology in real-world use. New insights in this paper include: empirical data showing the relationship between accuracy and the quality of model parameters; identification of correlations in speech recognition errors as a major cause of failures; and evidence for fundamental flaws in several components of current models. Taken together, these findings suggest several new research directions, and we briefly explore the potential of one of these.

In this paper, Sections 2 and 3 cover background material and the two dialog systems. Sections 4 and 5 then provide the analysis, and Section 6 concludes and suggests several new research directions.

2. STATISTICAL DIALOG SYSTEMS

Statistical dialog systems maintain a distribution over a set of hidden dialog states, such as the user's overall goal in the dialog or the user's true dialog act. For each dialog state s, a posterior probability of correctness called a *belief* is maintained $b(s)$. The set of hidden dialog states and their beliefs is collectively called the *belief state*, and updating the belief state is called *belief tracking*. Here we will present belief tracking at a level sufficient for our purposes; for a more general treatment, see [2].

At the start of the dialog, the belief state is initialized to a *prior* distribution $b_0(s)$. The system then takes an action a, and the user takes an action in response. The automatic speech recognizer and spoken language understanding (collectively called "ASR" in this paper) then produces a ranked list of N hypotheses for the user's action, $u_1, \ldots, u_N$, called an *N-best list*. For each N-best list the ASR also produces a distribution $P_{\text{asr}}(u)$ which assigns a local, context-independent probability of correctness to each item, often called a *confidence score*. The belief state is then updated:

$$b'(s) = k \cdot \sum_u P_{\text{asr}}(u) P_{\text{act}}(u|s,a) b(s) \qquad (1)$$

where $P_{\text{act}}(u|s,a)$ is the probability of the user taking action u given the dialog is in hidden state s and the system takes action a. k is a normalizing constant. In this paper, we'll assess whether the top belief state $s^* = \arg\max_s b(s)$ computed by Eq 1 yielded an improvement in accuracy compared to the top ASR result u_1 in two real-world dialog systems.

3. DIALOG SYSTEMS UNDER STUDY

The two systems under study in this paper – DS1 and DS2 – provide bus timetable information for Pittsburgh, USA. They were fielded to the public as a part of the Spoken Dialog Challenge [3]. They followed a highly directed flow, collecting one *slot* at a time. There are five slots: route, from, to, day, and time. These systems could only recognize values for the slot being queried, plus a handful of global commands ("repeat", "go back", "start over", "goodbye", etc.) – mixed initiative was not supported. The systems themselves were fielded by AT&T [4], and the analysis here is based on the

system recordings and logs, publically available from the Dialog Research Center at Carnegie Mellon University.

Each system opened by asking the user to say a bus route, or to say "I'm not sure." The systems could recognize any of the ∼100 routes in Pittsburgh, but could only provide times for a *covered* subset of routes. If an uncovered route was recognized, the system explained that it only had information for certain routes. Otherwise, the system next asked for the from and to slots. The system then asked if the caller wants times for the "next few buses". For the (few) callers who said "no", the system asked for the day then time in two separate questions. Finally bus times were read out. Users could say "start over" at any time.

Belief tracking was done with the AT&T Statistical Dialog Toolkit [5], and an independent belief state was maintained for each slot. After requesting the value of a slot, the system received an ASR N-best list, assigned each item a confidence score $P_{\text{asr}}(u)$, and updated the belief in (only) that slot using Eq 1. The top dialog hypothesis $s^* = \arg\max_s b(s)$ and its belief $b(s^*)$ were used to determine which action to take next, following a hand-crafted policy. This is in contrast to conventional dialog systems, in which the top ASR result governs dialog flow.

Confidence scores $P_{\text{asr}}(u)$ were assigned using a two-stage model [6]. In the first stage, a maximum entropy classifier assigned a probability to three classes, where the classes indicate (1) that the top ASR result u_1 is correct; (2) that one of the items in $u_2 \ldots u_N$ is correct; and (3) that none of the items on the ASR N-best list is correct. In the second stage, a Beta distribution is used to allocate the probability of class (2) across items $u_2 \ldots u_N$. The maximum entropy classifier and Beta distribution were trained on data (details in Section 4.1). The structure of the confidence score model $P_{\text{asr}}(u)$ made it possible for item $n = 2$ to be assigned a higher confidence score than $n = 1$, although this wasn't necessarily desired.

The two systems were nearly identical, except that DS1 could provide timetables for 8 covered routes and DS2 could provide timetables for ∼40 covered routes; DS2 used different priors b_0 than DS1; and DS2 used different training data to estimate $P_{\text{asr}}(u)$. Table 1 shows descriptive statistics of the dialogs.

Table 1: Two dialog systems studied in this paper. Utterance counts for each slot show the number of non-empty utterances received in response to system requests for that slot.

	DS1	DS2
Calls	779	1037
route utterances	1495	2955
from utterances	1197	1656
to utterances	1148	1592
day utterances	175	128
time utterances	155	237

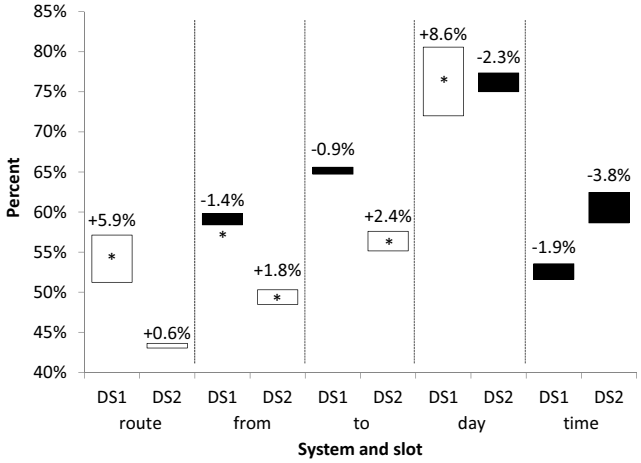


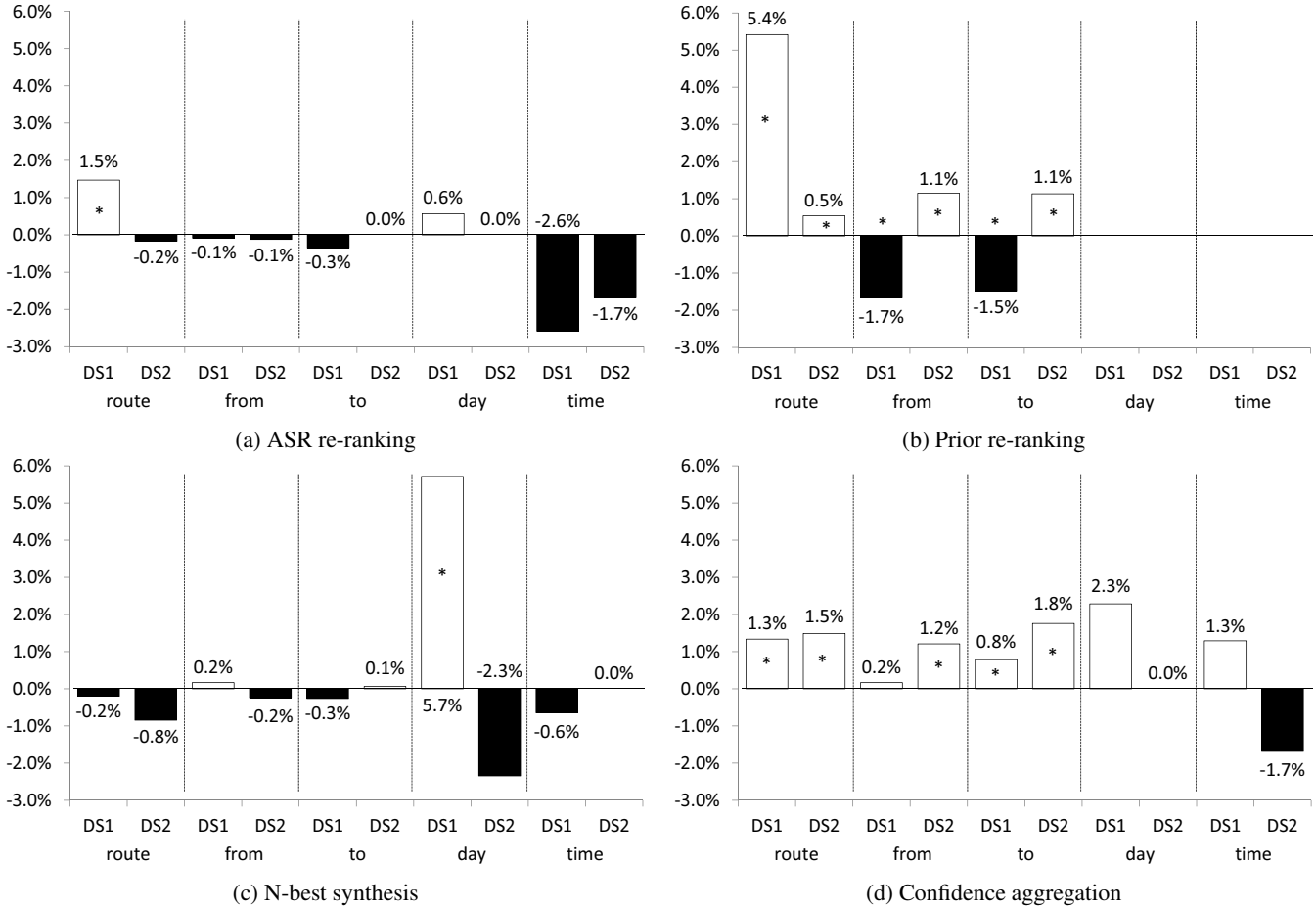
Fig. 1: Summary of accuracy. The tops and bottoms of each bar show accuracy for s^* and u_1. Unshaded bars indicate that the accuracy of s^* is higher than u_1 (ie, s^* corresponds to the top of the bar, and u_1 to the bottom). Shaded bars indicate that the accuracy of s^* is lower than u_1. Asterisk (*) indicates the difference is statistically significant with $p \leq 0.05$ using McNamara's Test.

4. EVALUATION OF ACCURACY

To measure the performance within each slot, we will compare the accuracy of the top belief state $s^* = \arg\max_s b(s)$ to the accuracy of the top ASR result u_1 (our baseline). We began by selecting utterances containing non-empty responses to each of the five slots (counts in Table 1). A professional transcriber (not the author) listened to each utterance, and marked each hypothesis on the ASR N-best list $u_1 \ldots u_N$ as *correct* if it was *semantically* consistent with the user's speech, or *incorrect* otherwise. Labels were checked by a second professional transcriber.

We next determined the accuracy of the top belief state s^*. In these systems, each item in the belief state maps directly to one or more ASR hypotheses. In addition, typically the user's goal remains fixed throughout the call, at least until the caller says "start over". Given this, the correctness of the top belief state was set to the correctness of the most recent ASR hypothesis it mapped to. However, if the user said "start over", the set of relevant ASR items was cleared. The accuracies for u_1 and s^* for each slot in each system are shown in Figure 1. While belief tracking yielded an improvement in accuracy in some cases, it caused a degradation in others.

We next sought to understand the causes of this varied performance. Formally, differences between the top ASR result u_1 and the top belief state s^* are simply the result of evaluating Eq 1. However, *intuitively* there are four *mechanisms*



(a) ASR re-ranking

(b) Prior re-ranking

(c) N-best synthesis

(d) Confidence aggregation

Fig. 2: Effects of each mechanism on each slot. Each bar shows $(x - y)/z$, where x is the number of utterances where the mechanism occurred *and* the belief 1-best is correct, y is the number of utterances where the mechanism occurred *and* the ASR 1-best is correct, and z is the total number of utterances in that slot/system (regardless of whether the mechanism occurred). Asterisk (*) indicates the difference is statistically significant with $p \leq 0.05$ using McNamara's Test.

which cause differences [1]:

- **ASR re-ranking**: Our confidence score $P_{\text{asr}}(u)$ had the ability to assign a higher confidence score to u_2 than u_1; when this *ASR re-ranking* happens, this may cause s^* to differ from u_1.
- **Prior re-ranking**: Statistical techniques use a prior probability for each possible dialog state – in our system, each slot value – $b_0(s)$. If an item recognized lower-down on the N-best list has a high prior, it can obtain the most belief, causing s^* to differ from u_1.
- **N-best synthesis**: If an item appears in two N-best lists, but is not in the top ASR N-best position in the latter recognition, it may still obtain the highest belief, causing s^* to differ from u_1.
- **Confidence aggregation**: If the top belief state s^* has high belief, then subsequent low-confidence recognitions which do not contain s^* will not dislodge s^* from the top position, causing s^* to differ from u_1.

Figure 2 shows the improvement/degradation of each mechanism on each slot/system. Although there are some trends, there is no overall pattern. The next four sections examine each mechanism in detail.

4.1. ASR re-ranking (Figure 2a)

Recall that the models that assigned (local) confidence scores P_{asr} could – as an artifact of their two-stage design – assign a higher confidence score to the $n = 2$ item than the $n = 1$ item. We call this re-ordering *ASR re-ranking*, and Figure 3 shows it consistently degraded ASR accuracy, with one exception (day in DS1).

DS1 and DS2 used different confidence models P_{asr}. When DS1 was launched, there was no same-system data available, so a large corpus of data from a different dialog

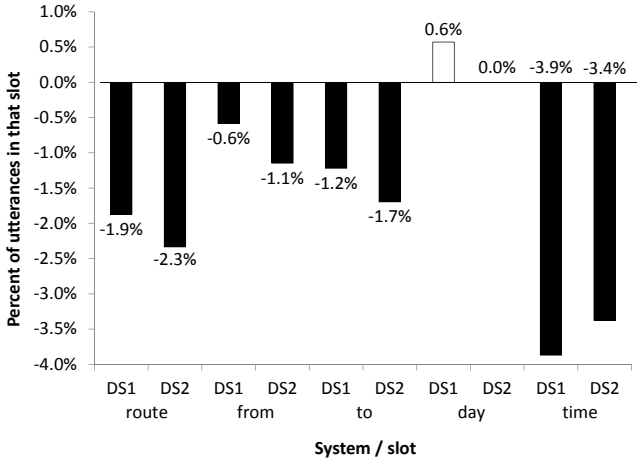


Fig. 3: Effect of ASR re-ranking on local accuracy. Bars show $(x - y)/z$, where x is the number of correct u^* where $u^* = \arg\max_u P_{\text{asr}}(u)$, y is the number of correct u_1, and z is the total number of utterances in the system/slot.

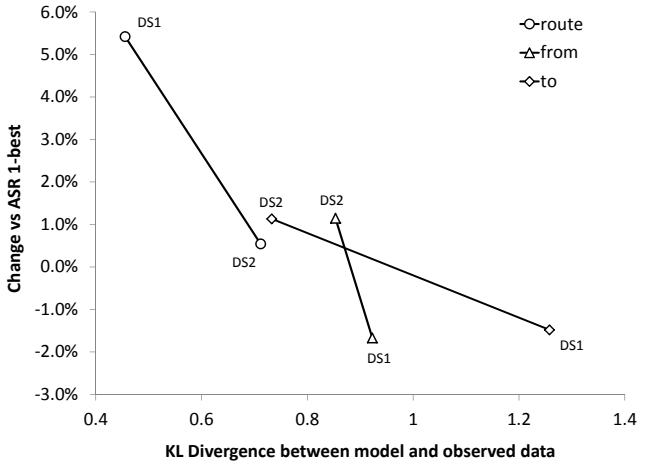


Fig. 4: Discrete KL divergence between model prior and observed data vs. the change in accuracy compared to the ASR 1-best. The y axis is computed as in Figure 2. Increasing KL divergence (i.e., poorer model fit) degrades the accuracy of belief tracking.

system was used to train the models [7]. This mismatch was one possible cause of the degradation for DS1, so P_{asr} for DS2 was trained on data from DS1. However, as shown in Figure 3, ASR re-ranking also reduced ASR accuracy in DS2. This suggests that mis-matched training data is not the primary cause. Rather, it seems a more sophisticated model for P_{asr} is required – i.e., one which is explicitly aware of the order of items on the N-best list.

4.2. Prior re-ranking (Figure 2b)

Non-uniform priors were used in only route, from, and to. Figure 2b shows that prior re-ranking improved accuracy for route, substantially for DS1 and marginally for DS2. It also improved accuracy for from and to in DS2, but degraded accuracy for these slots in DS1. The explanations for these results lay in key differences between DS1 and DS2.

The first key difference between DS1 and DS2 is how priors were estimated. In DS1, an attempt was made to estimate priors using a heuristic that avoided collecting usage data. The heuristic assigned a prior proportional to the *number of bus stops* the slot value referred to. For example, for locations (from and to), "downtown" referred to many bus stops, but "the airport" referred to just one. In DS2, priors were estimated from actual usage observed in DS1.

For locations in DS1, this heuristic was a failure. The problem is that the heuristic did not reflect the fact that certain stops are more *popular* than others: for example, the airport corresponded to a single bus stop, but it was very popular. The net effect was that prior re-ranking for locations in DS1 degraded performance. In DS2, with priors estimated from (transcribed) usage data rather than a heuristic, priors yielded an improvement in accuracy for locations.

The second key difference is that DS2 covered many more bus routes than DS1. Most requests were for covered routes, which had high priors; uncovered routes had very low priors. In DS1, the result was that most recognitions of non-covered routes were errors; the strong prior moved covered routes to the top of the belief state, yielding a large improvement for belief tracking for route in DS1. In DS2, a larger set of routes were covered, so erroneous recognitions were no longer obvious. As a result, prior re-ranking still helped for route in DS2, but to a lesser extent.

The overall trend is that the effectiveness of prior re-ranking depends on how well the prior matches real use. Figure 4 shows the discrete Kullback-Leibler (KL) divergence between the frequency of observation and the prior used in deployment for each of these 3 slots across the 2 systems. Within each slot, as the KL divergence increases, accuracy of belief tracking decreases.

4.3. N-best synthesis (Figure 2c)

Performance for N-best synthesis was quite varied. For route, from, to, and time, there was generally a negative (or marginal) effect. For day, there was a large improvement for DS1, and a moderate degradation for DS2.

We manually examined each instance of a degradation and found that 86% of failed instances of N-best synthesis were caused by *correlated ASR errors*: i.e., the same recognition error occurring repeatedly. Figure 5 shows an illustration of ASR error correlation. The key problem is that the update in Eq 1 – in particular P_{asr} – assumes that confusions are independent. Correlations cause repeated errors to be wrongly assigned too much belief mass.

Looking at day in DS2, we found a secondary cause for

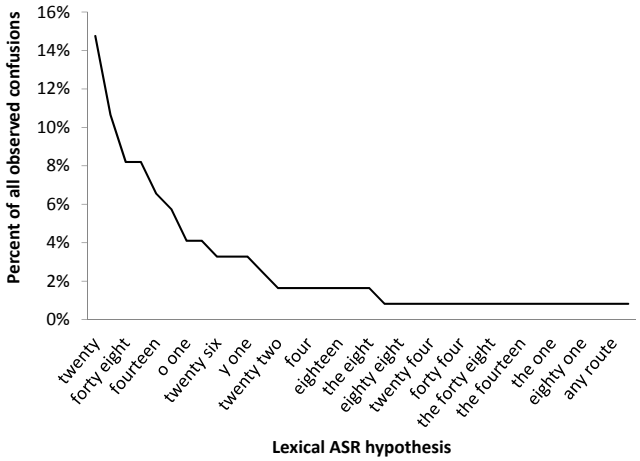


Fig. 5: Semantically incorrect items appearing in any location on the ASR N-best list when the user said *twenty eight x*. For space on the x axis, every second item is shown. The skew of the curve shows that confusions are highly correlated.

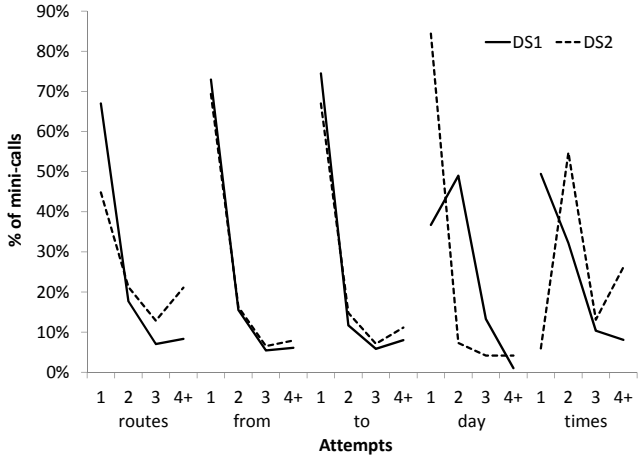


Fig. 6: Histogram of number of times each slot was requested by the system. The y axis shows the percent of each mini-call, where a mini-call is the same as a call except that "start over" begins a new mini-call.

degradations. Here, most of the degradations were caused by the user saying "no" in response to the system confirming the *correct* item, even though the user subsequently asked for the same item again. The user behavior model P_{act} – which was based on hand-crafted heuristics – assigned a zero probability to this seemingly irrational behavior. As a result the correct item was ranked very low in the belief state.

Listening to these calls revealed that the confirmation wording for day was creating confusion. For example, for a call on Friday, one user said "today" but the system asked "Did you say Friday?". In addition to improving this confirmation strategy, it is clear that the user action model (like the priors) can be difficult to predict and should be estimated from real usage data.

4.4. Confidence aggregation (Figure 2d)

Figure 2d shows that confidence aggregation had an overall positive effect, with day in DS1 being the most pronounced. The one exception was time in DS2, where there was a negative effect. Confidence aggregation has more opportunity to occur when questions are more often asked repeatedly. Figure 6 shows histograms of how many times each slot was requested. In most cases, slots were most often requested once; however, day in DS1 and time in DS2 were usually requested more times.

Based on past investigation, we were aware that day in DS1 had a bug that set priors to be an order of magnitude too low [1]. As a result, more requests were required to obtain belief values above the (manually-set) threshold required to progress. This bug in day in DS1 was fixed in DS2. Unfortunately we found that the same problem was inadvertently introduced to time in DS2. Thus these questions were more often asked repeatedly, illustrated by the disproportionately high counts of day utterances in DS1 and time utterances in DS2 in Table 1.

But why was belief tracking accuracy for day in DS1 improved, whereas time in DS2 was degraded? The underlying cause was ASR re-ranking errors earlier in the dialogs. For day in DS1, ASR re-ranking yielded a small (anomalous) improvement to ASR accuracy; for time in DS2, ASR re-ranking yielded a large degradation to ASR accuracy (Figure 3). Confidence aggregation amplifies these effects by carrying them forward in the dialog.

5. EVALUATION OF DISCRIMINATION

The analysis in the preceding sections assessed the *accuracy* of the top hypothesis in the belief state. In practice, a system must decide whether to accept or reject a hypothesis, so it is also important to evaluate the ability of the belief state to *discriminate* between correct and incorrect hypotheses. We studied this by plotting receiver operating characteristic (ROC) curves for each slot, in Figure 7. The ASR 1-best u_1 is shown using the computed $P_{\text{asr}}(u_1)$, and the belief 1-best s^* is shown using its belief $b(s^*)$.

Where the belief state has higher accuracy – route and day in DS1 – the belief state shows better ROC results, especially at higher false-accept rates. However, gains in ROC performance appear to be due entirely to gains in accuracy: in slots where accuracy is similar between belief tracking and ASR, the belief state shows similar or worse performance. time in DS2 was particularly affected, by the negative effect of ASR re-ranking, further compounded by confidence aggregation. Overall, the trend appears to be that if belief tracking does not improve over ASR 1-best, then belief tracking does not enable better accept/reject decision to be made.

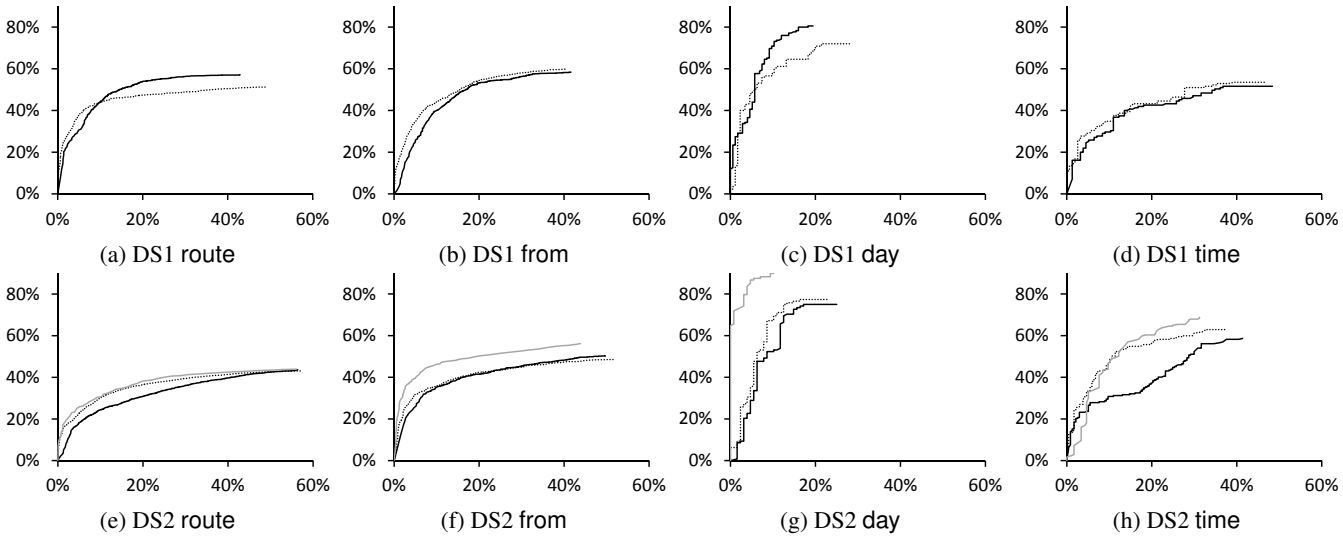


Fig. 7: ROC curves. to is very similar to from and is omitted for space. X-axis shows false accepts, Y-axis shows true accepts. The solid black line is the belief 1-best; the dotted line is the ASR 1-best. The difference in each pair of curves' maximum values on the Y-axis corresponds to bar heights in Figure 1. In the lower panels, the gray line shows results from a discriminative classifier trained on data from DS1, and described in Section 6.

6. CONCLUSION AND FUTURE DIRECTIONS

This paper has presented an analysis of 2 versions of one of the first statistical dialog systems in public use. Overall, the findings have underscored the importance (and difficulty!) of correctly estimating each model component. Mismatches in all 3 component models – i.e., the models of ASR errors P_{asr}, user behavior P_{act}, and goal priors b_0 – caused degradations compared to the top speech recognition hypothesis.

More fundamentally, the analysis here suggests crucial weaknesses in the formulation of the model, not merely in parameter estimates. For example, ASR error correlations are not currently being modeled, and they are harming performance. The lackluster discrimination in the belief state is more troubling, suggesting that the formulation of the update as a generative model (Eq 1) may be problematic. Discriminative models for dialog tracking – which are trained directly on the data and explicitly optimized for discrimination – are a natural alternative [8]. To briefly highlight their potential, we identified about 60 features, configured slot-specific discriminative classifiers to predict $b(s)$, trained on data from DS1, and tested on data from DS2. Results are included in Figures 7e-7h. In most cases the discriminative method attains both higher accuracy (a larger maximum value on y-axis) and better discrimination than both the ASR 1-best and generative belief state. To obtain a reliable comparison, the discriminative models should also be tested in a public deployment; however this preliminary result does suggest there is substantial room for improvement over current methods.

Despite the issues identified in this paper, the first public deployments have nonetheless shown that – when models are properly estimated – statistical approaches can indeed achieve their aim of increasing robustness to ASR errors.

7. REFERENCES

[1] JD Williams, "An Empirical Evaluation of a Statistical Dialog System in Public Use," in *SIGdial*, 2011.

[2] JD Williams and SJ Young, "Partially observable Markov decision processes for spoken dialog systems," *Computer Speech and Language*, vol. 21, no. 2, pp. 393–422, 2007.

[3] AW Black et al, "Spoken dialog challenge 2010: Comparison of live and control test results," in *SIGdial*, 2011.

[4] JD Williams, I Arizmendi, and AD Conkie, "Demonstration of AT&T "Let's Go": A production-grade statistical spoken dialog system," in *Proc SLT*, 2010.

[5] "AT&T Statistical Dialog Toolkit," `www2.research.att.com/sw/tools/asdt/`.

[6] JD Williams and S Balakrishnan, "Estimating probability of correctness for ASR N-Best lists," in *SIGdial*, 2009.

[7] G Parent and M Eskenazi, "Toward Better Crowdsourced Transcription: Transcription of a Year of the Let's Go Bus Information System Data," in *Proc SLT*, 2010.

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