An Empirical Evaluation of a Statistical Dialog System in Public Use

Jason D. Williams
AT&T Labs - Research, Shannon Laboratory, 180 Park Ave., Florham Park, NJ 07932, USA
jdw@research.att.com

Abstract

This paper provides a first assessment of a statistical dialog system in public use. In our dialog system there are four main recognition tasks, or slots – bus route names, bus-stop locations, dates, and times. Whereas a conventional system tracks a single value for each slot – i.e., the speech recognizer's top hypothesis – our statistical system tracks a distribution of many possible values over each slot. Past work in lab studies has showed that this distribution improves robustness to speech recognition errors; but to our surprise, we found the distribution yielded an increase in accuracy for only two of the four slots, and actually decreased accuracy in the other two. In this paper, we identify root causes for these differences in performance, including intrinsic properties of N-best lists, parameter settings, and the quality of statistical models. We synthesize our findings into a set of guidelines which aim to assist researchers and practitioners employing statistical techniques in future dialog systems.

1 Introduction

Over the past decade, researchers have worked to apply statistical techniques to spoken dialog systems, and in controlled laboratory studies, statistical dialog systems have been shown to improve robustness to errors compared to conventional approaches (Henderson and Lemon, 2008; Young et al., 2010; Thomson and Young, 2010). However, statistical techniques have not yet been evaluated in a publicly deployed system, and real users behave very differently to usability subjects (Raux et al., 2005; Ai et al., 2008). So there is an important open question whether statistical dialog systems improve performance *with real users*.

This paper provides a first evaluation of a publically deployed statistical dialog system, AT&T Let's Go (Williams et al., 2010). AT&T Let's Go provides bus times for Pittsburgh, and received approximately 750 calls from real bus riders during the 2010 Spoken Dialog Challenge (Black et al., 2010). AT&T Let's Go is based on a publicly available toolkit (Williams, 2010a) and achieved the highest rates of successful task completion on real callers in the challenge, so it provides a relevant exercise from which to draw inferences.

AT&T Let's Go collected four types of information, or *slots*: bus route names, bus-stop names, dates, and times. For each slot, we measured turn-level accuracy of the deployed statistical system and compared it to accuracy without application of the statistical techniques (i.e., the top speech recognition result).

To our surprise, we found that statistical techniques appeared to improve accuracy for only two of the four slots, and decreased accuracy for the other two. To investigate this, we considered four *mechanisms* by which statistical methods can differ from the top speech recognition result. Analyzing the effects of each mechanism on each slot enables underlying causes to be identified: for example, one mechanism performed exceptionally well when its statistical models was well matched to usage data, but rather poorly when its model diverged from real usage. We believe this analysis – the focus of this paper – is relevant to researchers as well as practi-

tioners applying statistical techniques to production systems.

In this paper, Section 2 reviews the operation of statistical spoken dialog systems. Section 3 then describes the AT&T Let's Go dialog system. Section 4 reports on overall accuracy, then analyzes the underlying reasons for accuracy gains and losses. Section 5 tackles how well error in the belief state can be *identified* compared to speech recognition errors. Section 6 concludes by summarizing lessons learned.

2 Statistical dialog systems

Statistical dialog systems maintain a distribution over a set of hidden dialog states. A dialog state includes information not directly observable to the dialog system, such as the user's overall goal in the dialog or the user's true action (e.g., 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 (Williams and Young, 2007).

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 (ASR) then produces a ranked list of N hypotheses for the user's action, $\mathbf{u} = (u_1, \ldots, u_N)$, 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 practice specialized techniques must be used to compute Eq 1 in real-time. The system in this paper uses *incremental partition recombination* (Williams, 2010b); alternatives include the Hidden Information State (Young et al., 2010), Bayesian Update of Dialog States (Thomson and Young, 2010), and particle filters (Williams, 2007). The details are not important for this paper – the key idea is that Eq 1 synthesizes a prior distribution over dialog states together with all of the ASR N-best lists and local confidence scores to form a cumulative, whole-dialog posterior probability distribution over all possible dialog states, $b(s)$.

In the system studied in this paper, slots are queried separately, and an independent belief state is maintained for each. Consequently, within each slot user actions u and hidden states s are drawn from the same set of slot values. Thus the top ASR result u_1 represents the ASR's best hypothesis for the slot value in the current utterance, whereas the top dialog state $\arg\max_s b(s) = s^*$ represents the belief state's best hypothesis for the slot value given all of the ASR results so far, a prior over the slot values, and models of user action likelihoods. The promise of statistical dialog systems is that s^* will (we hope!) be correct more often than u_1. In the next section, we measure this in real dialogs.

3 AT&T Let's Go

AT&T Let's Go is a statistical dialog system that provides bus timetable information for Pittsburgh, USA. This system was created to demonstrate a production-grade system built following practices common in industry, but which incorporates two statistical techniques: belief tracking with the AT&T Statistical Dialog Toolkit (Williams, 2010a), and regression-based ASR confidence scores (Williams and Balakrishnan, 2009).

As with most commercial dialog systems, AT&T Let's Go follows a highly directed flow, collecting one *slot* at a time. There are four types of slots: ROUTE, LOCATION, DATE, and TIME. The system can only recognize values for the slot being queried, plus a handful of global commands ("repeat", "go back", "start over", "goodbye", etc.) – mixed initiative and over-completion were not supported. As mentioned above, an independent belief state is maintained for each slot: this was an intentional design decision made in order to use statistical techniques within current commercial practices.

The system opens by asking the user to say a bus ROUTE, or to say "I'm not sure." The system next asks for the origin and destination LOCATIONs. The system then asks if the caller wants times for the "next few buses"; if not, the system asks for the DATE then TIME in two separate questions. Finally bus times are read out.

After requesting the value of a slot, the system receives an N-best list, assigns each item a confidence score $P_{\text{asr}}(u)$, and updates the belief in (only) that slot using Eq 1. The top dialog hypothesis s^* and its belief $b(s^*)$ are used to determine which action to take next, following a hand-crafted policy. This is in contrast to a conventional dialog system, in which the top ASR result and its confidence govern dialog flow. Figure 6 shows the design of AT&T Let's Go.

In the period July 16 – August 16 2010, AT&T Let's Go received 742 calls, of which 670 had one or more user utterances. These calls contained a total of 8269 user utterances, of which 4085 were in response to requests for one of the four slots. (The remainder were responses to yes/no questions, timetable navigation commands like "next bus", etc.)

Our goal in this paper is to determine whether tracking a distribution over multiple dialog states improved turn-level accuracy compared to the top ASR result. To measure this, we compare the accuracy of the top belief state and the top ASR result. A transcriber listened to each utterance and marked the top ASR hypothesis as *correct* if it was an exact lexical or semantic match, or *incorrect* otherwise. The same was then done for the top dialog hypothesis in each turn.

Accuracy of the top ASR hypothesis and the top belief state are shown in Table 1, which indicates that belief monitoring improved accuracy for ROUTE and DATE, but degraded accuracy for LOCATION and TIME. We had hoped that belief tracking would improve accuracy for all slots; seeing that it hadn't prompted us to investigate the underlying causes.

Slot	ROUTE	LOCATION	DATE	TIME
Utts	1520	2235	173	157
ASR	769	1326	124	80
correct	50.6%	59.3%	71.7%	51.0 %
Belief	799	1246	139	63
correct	52.6%	55.7%	80.3%	40.1%
Belief	+30	-80	+15	-17
– ASR	+2.0%	-3.6%	+8.7%	-10.8%

Table 1: Accuracy of the top ASR result and top belief state. LOCATION includes both origin and destination utterances. Most callers requested the next bus so few were asked for DATE and TIME.

4 Belief tracking analysis

When an ASR result is provided to Eq 1 and a new belief state is computed, the top dialog state hypothesis s^* may differ from top ASR result u_1. Formally, these differences are simply the result of evaluating this equation. However, *intuitively* there are four *mechanisms* which cause differences, and each difference can be explained by the action of one or more mechanisms. These mechanisms are summarized here; the appendix provides graphical illustrations.[1]

- **ASR re-ranking**: When computing a confidence score $P_{\text{asr}}(u)$, it is possible that the entry with the highest confidence $u^* = \arg\max_u P_{\text{asr}}(u)$ will not be the first ASR result, $u_1 \neq u^*$. In other words, if the confidence score *re-ranks* the N-best list, this may cause s^* to differ from u_1 (Figure 7).
- **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 (Figure 8).
- **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 9).
- **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 (Figure 10).

[1]This taxonomy was developed for belief tracking over a single slot. For systems which track joint beliefs over multiple slots, additional mechanisms could be identified.

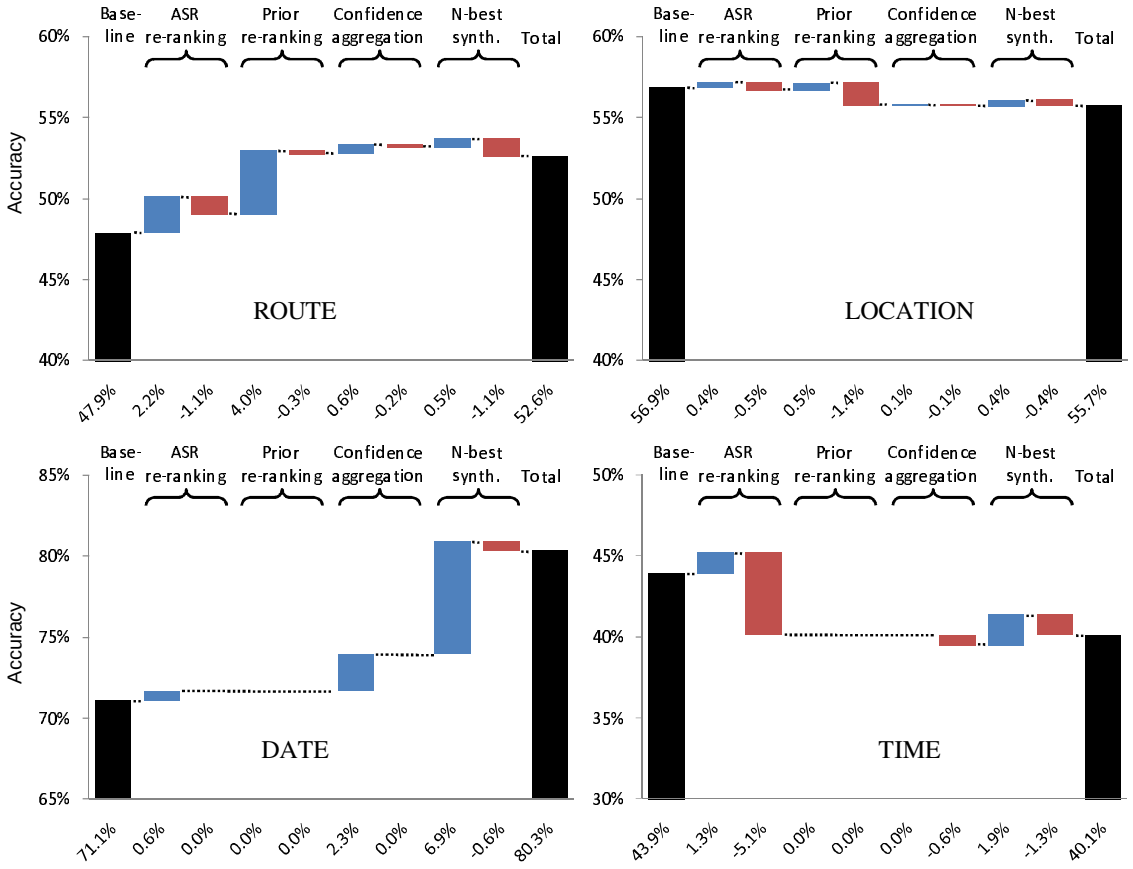

Figure 1: Differences in accuracy between ASR and belief monitoring. "Baseline" indicates accuracy among utterances where belief monitoring had no effect – where ASR and belief monitoring are both correct, or both incorrect. Blue bars show cases where the top belief state s^* is correct and the top ASR result u_1 is not; red bars show cases where u_1 is correct and s^* is not. The plot is arranged to show a running total where blue bars increase the total and red bars decrease the total. Percentages under blue and red bars show the change in accuracy due to each mechanism. The black bar on the right shows the resulting accuracy in deployment.

We selected utterances where the correctness of the top ASR result and top dialog hypothesis differed – where one was correct and the other was not – and labeled these by hand to indicate which of the four mechanisms was responsible for the difference. In a few cases multiple mechanisms were responsible; these were labeled with the first contributing mechanism in the order listed above.

Figures 1 shows results. Of the four mechanisms, prior re-ranking occurred most often, and confidence aggregation occurred least often. Interestingly, some mechanisms provided a performance gain for certain slots and a degradation for others. This led us to look at each mechanism in detail.

4.1 Evaluation of ASR Re-ranking

The recognizer used by AT&T Let's Go produced an N-best list ordered by decoder cost. After decoding, a confidence score was assigned to each item on the N-best list using a regression model that operated on features of the recognition (Williams and Balakrishnan, 2009). The purpose of this regression was to assign a probability of correctness to each item on the N-best list; while it was not designed to re-rank the N-best list, the design of this model did allow it to assign a higher score to the $n = 2$ hypothesis than the $n = 1$ hypothesis. When this happens, we say the N-best list was *re-ranked*. Table 2 shows how often ASR re-ranking occurred, and how often the

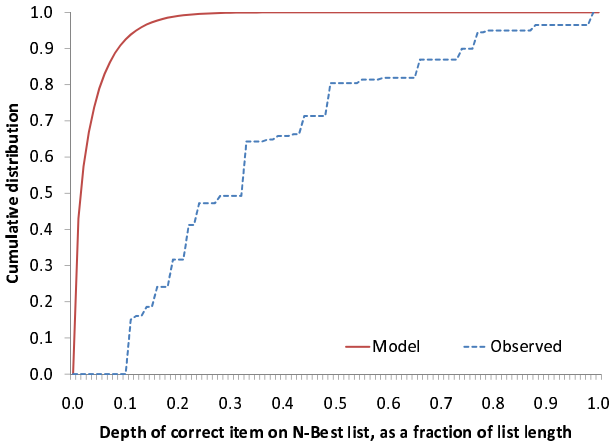

Figure 2: Cumulative distribution of the position of the correct item on N-Best lists for the ROUTE when the correct item is in position $2 \ldots N$. Depth is shown as a fraction of the N-Best list length.

ASR re-ranking helped and hurt ASR accuracy. We found that re-ranking degraded ASR accuracy for all slots, except DATE where it had a trivial positive impact. This suggested a problem with our confidence score; examining ROUTE, LOCATION, and TIME we found that the distributions used by the confidence score that apportions mass to items $2 \ldots N$ were far more concentrated on the N=2 entry than observed in deployment (Figure 2). Investigation revealed a bug in the model estimation code for these slots.

Where ASR re-ranking decreased ASR accuracy, we'd expect to see it also decrease belief state accuracy. Indeed, for the TIME slot, ASR re-ranking causes a substantial decrease in belief state accuracy, highlighting the importance of an accurate confidence score to statistical techniques. However, for the ROUTE slot, we see an *increase* in belief state accuracy attributed to ASR re-ranking. This can be explained by interaction between ASR re-ranking and prior re-ranking, discussed next.

4.2 Evaluation of prior re-ranking

Whereas N-best re-ranking affects $b'(s)$ via P_{asr}, *prior re-ranking* affects $b'(s)$ via the *prior probability* in a slot $b_0(s)$ – i.e., the initial belief, at the start of the dialog, for each value the slot may take. If the slot's prior is uniform (non-informative), we expect to see no effect on accuracy due to the prior – indeed, Figure 1 shows that priors had no effect on belief accuracy for DATE and TIME, which used uniform priors.

ROUTE and LOCATION employed a non-uniform prior, and here we'd expect to see a gain in performance if the prior matches actual use. Both priors were computed using a simple heuristic in which the prior was proportional to the number of distinct bus-stops on the route or covered by the location expression, smoothed with a smoothing factor. For example, the phrase "downtown" covered 17 stops and its prior was 0.018; the phrase "airport" covered 1 stop and its prior was 0.00079. Even though historical usage data was available to Spoken Dialog Challenge 2010 participants (Parent and Eskenazi, 2010), we instead chose to base priors on bus-stop counts as a test of whether effective priors could be constructed without access to usage data.

Overall the prior for ROUTE fit actual usage data well (Figure 3), and we see a corresponding net gain in belief accuracy of $3.7\% = 4.0\% - 0.3\%$ in Figure 1. However the prior for LOCATION was a poor match with actual usage (Figure 4), and this caused a net degradation in belief accuracy of $-0.9\% = 0.5\% - 1.4\%$. The key problem is that the heuristic wrongly assumed all stops are equally popular: for example, although the airport contained a single stop (and thus received a very low prior), it was very popular. This suggests that it would be better to estimate priors based on usage data rather than the bus-stop count heuristic. More broadly, it also underscores the importance of accurate priors to statistical dialog techniques.

In the previous section, for ROUTE, it was observed that ASR re-ranking degraded ASR accuracy, yet caused an improvement in belief accuracy. The effects of the prior explain this: the prior was often stronger, such that an error introduced by ASR re-ranking was cancelled by prior re-ranking. Examining cases where ASR re-ranking occurred but the belief state was still correct confirmed this. Where ASR re-ranking and prior re-ranking agreed, the ASR re-ranking received credit. Looking at LOCATION, the prior was essentially noise, so ASR re-ranking errors could not be systematically canceled by prior re-ranking in the same way – indeed, LOCATION belief accuracy was degraded by both ASR re-ranking and prior re-ranking. More broadly, this provides a nice illustration of how statistical tech-

Slot	ROUTE	LOCATION	DATE	TIME
All utterances	1520	2235	173	157
Utterances with ASR re-ranking	505 33.2%	305 13.6%	3 1.7%	40 25.5%
ASR re-ranked; N=2 correct (ASR re-ranking helped)	36 +2.4%	11 +0.5 %	1 +0.6 %	3 +1.9 %
ASR re-ranked; N=1 correct (ASR re-ranking hurt)	63 -4.1%	33 -1.5 %	0 0 %	9 -5.7 %
Net gain from ASR re-ranking	-27 -1.8 %	-22 -1.0%	+1 +0.6%	-6 -3.8%

Table 2: ASR re-ranking.

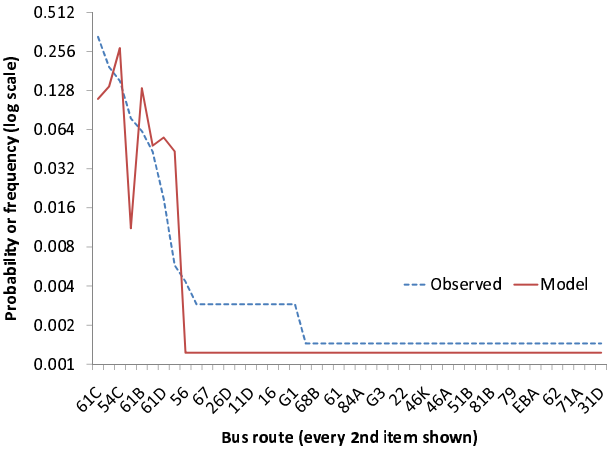

Figure 3: Modeled prior for ROUTE vs. observed usage. The modeled prior was a relatively good predictor of actual usage.

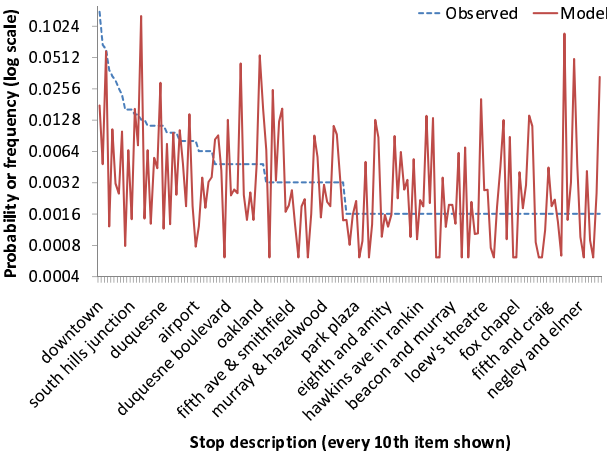

Figure 4: Modeled prior for LOCATION vs. observed usage. The modeled prior was essentially noise compared to actual usage.

niques can combine conflicting evidence – in this case, from the prior and ASR.

4.3 Evaluation of confidence score aggregation

The conditions for confidence score aggregation occur somewhat rarely: for no slot did it have the greatest effect on belief accuracy. It had the largest effect on DATE; investigation revealed that belief scores for DATE were relatively lower than for other slots (Table 3). Since all slots used the same thresholds to make accept/reject decisions, DATE had proportionally more retries in which the top belief hypothesis was correct, yielding more opportunities for confidence aggregation to have an effect.

But why were belief values for DATE lower than for other slots? Investigation revealed that a bug

Slot	ROUTE	LOCATION	DATE	TIME
Correct	0.90	0.89	0.60	0.73
Incorrect	0.52	0.59	0.34	0.53

Table 3: Average belief in the top dialog state hypothesis when that hypothesis was correct or incorrect.

was causing priors for DATE to be nearly an order of magnitude too small, so that each recognized date was artificially improbable. As a result, DATE effectively had a more stringent threshold for accept/reject decisions. Although caused by a bug, this case study provides a more general illustration: obtaining sufficient belief to meet higher thresholds requires more ASR evidence in the form of more re-

Slot	ROUTE	LOCATION	DATE	TIME
Average N-best list length	5.0	2.8	2.1	4.3
N-best accuracy	27.9%	10.6%	46.0%	34.7%
Average position of correct item ($n > 1$)	3.3	3.2	2.6	2.9

Table 4: Descriptive statistics for N-best lists. *Average N-best list length* indicates the average length of all N-best lists, regardless of accuracy. *N-best accuracy* indicates how often the correct item appeared in any position $n > 1$ among cases where the top ASR result $n = 1$ was not correct. *Average position of correct item* refers to the average n among cases where the correct item appeared with $n > 1$.

tries.

4.4 Evaluation of N-best synthesis

For DATE, N-best synthesis had a large positive effect, TIME and LOCATION a small positive effect (or no effect), and ROUTE a small negative effect. N-best synthesis occurs when commonality exists across N-best lists, so we next examined the N-best lists for each slot.

Table 4 shows three key properties of the N-best lists. ROUTE and DATE had the most extreme values: ROUTE had the longest N-best lists, comparatively poor N-best accuracy, and the correct item appeared furthest down the N-best list. By contrast, DATE had the shortest N-best lists, the best N-best accuracy, and the correct item appeared closest to the top. LOCATION and TIME were between the two. This relative ordering aligns with the observed effect that N-best synthesis had on belief accuracy, where DATE enjoyed a large improvement and ROUTE suffered a small degradation.

This correlation suggests that basic properties of the N-best list govern the effectiveness of N-best synthesis: when N-best lists are shorter, more often contain the correct answer, and when the correct answer is closer to the top position, N-best synthesis can lead to large gains. When N-best lists are longer, less often contain the correct answer, and when the correct answer is farther from the top position, N-best synthesis can lead to small gains or even degradations.

5 Identifying belief state errors

The analysis in the preceding section assessed the *accuracy* of 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 5.

Where the belief state has higher accuracy (ROUTE, DATE), the belief state shows somewhat better ROC results, especially at higher false-accept rates. However, gains in ROC performance appear to be due entirely to gains in accuracy: In LOCATION, belief tracking made nearly no difference to accuracy, and the belief state shows virtually no difference to ASR in ROC performance. TIME suffered degradations in both accuracy and ROC performance. The trend appears to be that if belief tracking does not improve over ASR 1-best, then it seems that belief tracking does not enable better accept/reject decision to made. Perhaps addressing the model deficiencies mentioned above will improve discrimination – this is left to future work.

6 Conclusions

This paper has provided a first assessment of statistical techniques in a spoken dialog system under real use. We have found that belief tracking is not guaranteed to improve accuracy – its effects vary depending on the operating conditions:

- Overall the effects of prior re-ranking and N-best synthesis are largest; confidence aggregation has the smallest effect.
- When N-best lists are *useful*, N-best synthesis can have a large positive effect (DATE); when N-best lists are more noisy, N-best synthesis has a small or even negative effect (ROUTE).
- In the presence of more rejection, confidence aggregation can have a positive effect (DATE),

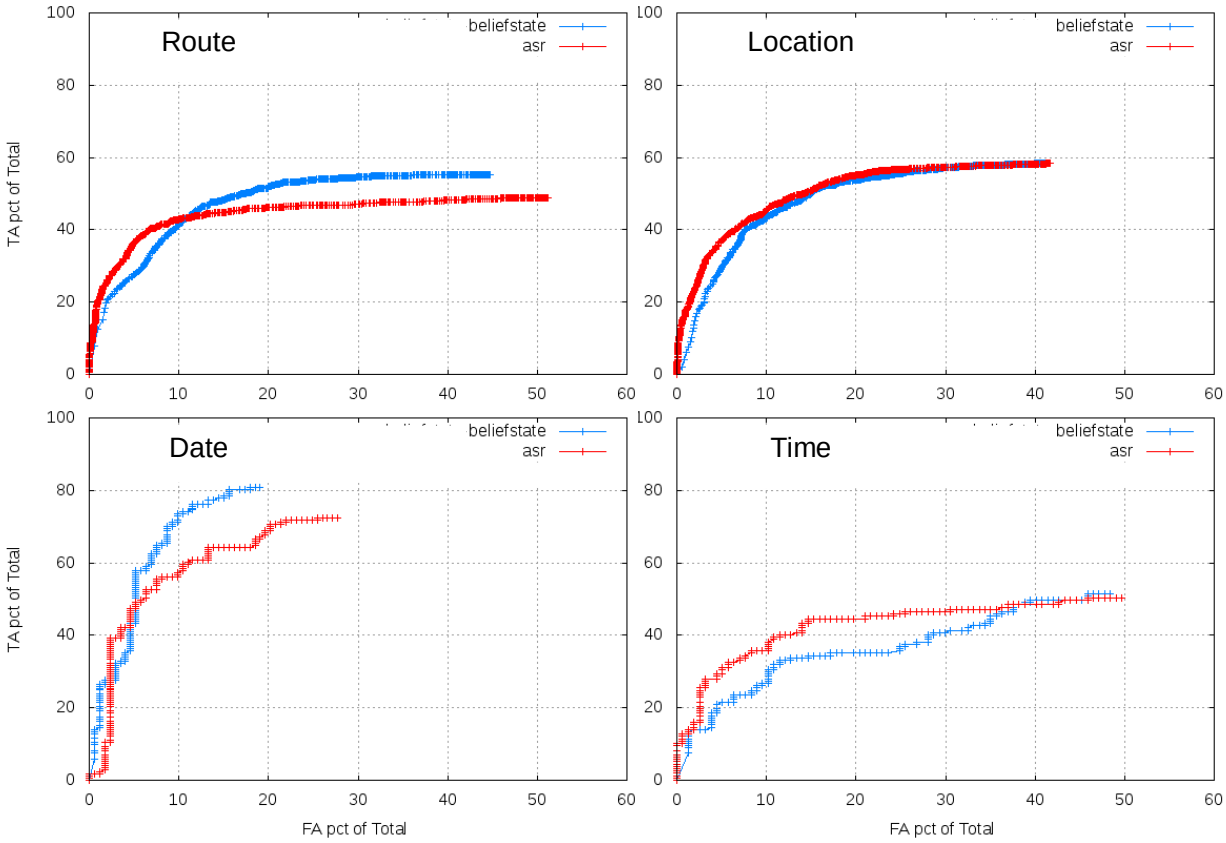

Figure 5: ROC curves. Red curves show the top-scored ASR hypothesis u^* with accept/reject decisions made using the confidence score $P_{\text{asr}}(u)$; blue curves show the top belief state s^* with accept/reject decisions made using its belief $b(s^*)$.

but otherwise plays a small role.

- When there exists an informative prior and it is estimated correctly, prior re-ranking produces an accuracy gain (ROUTE); when estimated poorly, it degrades accuracy (LOCATION).
- The belief state, at least when using our current models, improves accept/reject decisions only when belief tracking produces a gain in accuracy over ASR. Absent an accuracy increase, the belief state is no more informative than a good confidence score for making accept/reject decisions.

We believe these findings validate that statistical techniques – properly employed – have the capability to improve ASR robustness under real use. This paper has focused on descriptive results; in future work, we plan to test whether correcting the model deficiencies and re-running belief tracking does indeed improve performance. For now, we hope that this work serves as a guide to practitioners building statistical dialog systems, providing some instruction on the importance of accurate model building, and examples of the effects of different design decisions.

Acknowledgments

Thanks to Barbara Hollister and the AT&T labeling lab for their excellent work on this project.

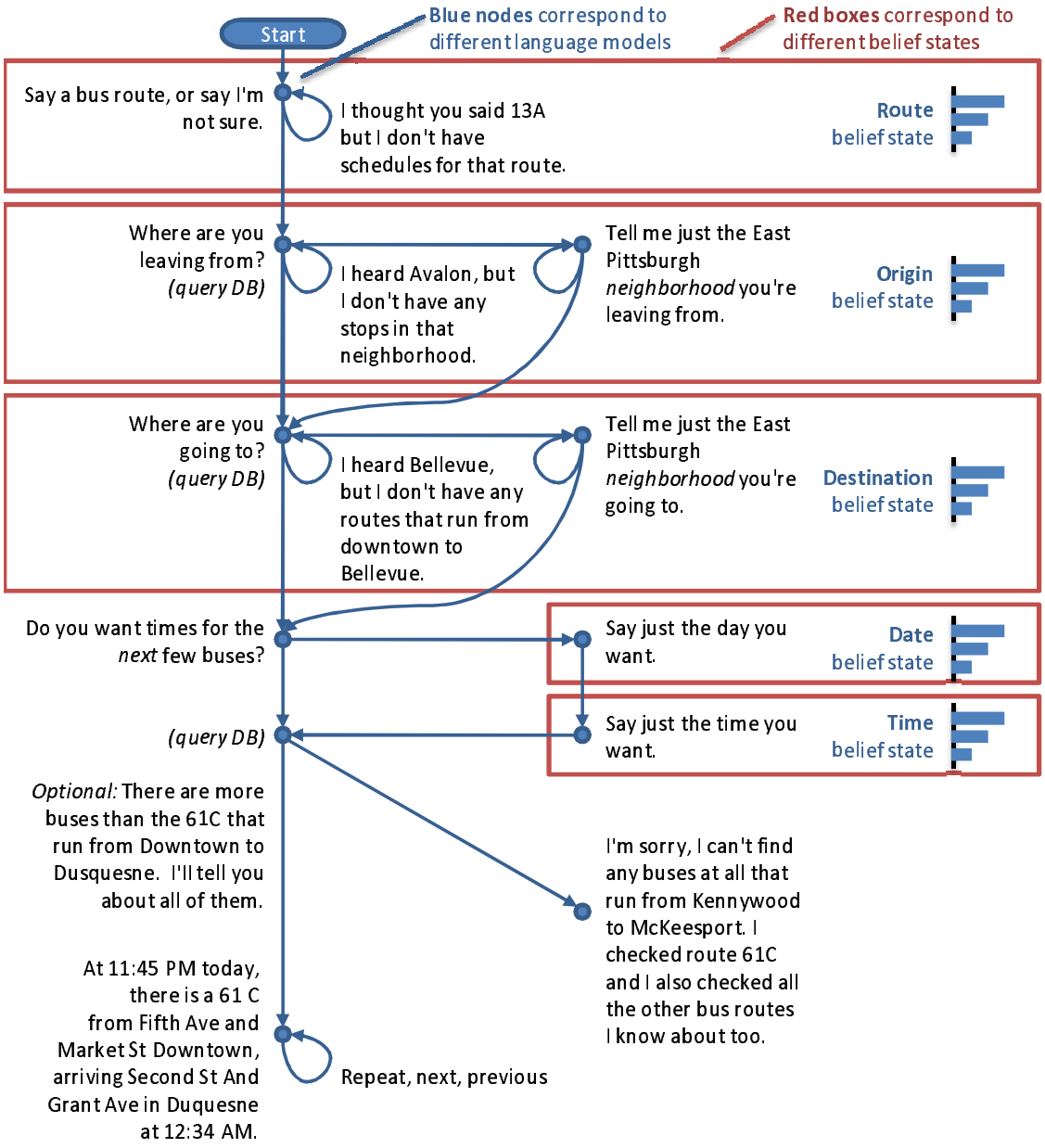

Figure 6: Flowchart of AT&T Let's Go. The system asks for the bus route, then the origin bus stop, then the destination bus stop. If the user does not want the next few buses, the system also asks for the date and time. Prompts shown are paraphrases; actual system prompts include example responses and are tailored to dialog context. Different language models are used for each slot, and separate belief states are maintained over each of these 5 slots. In the analysis in this paper, results for the origin and destination slots have been combined to form the LOCATION slot.

References

H Ai, A Raux, D Bohus, M Eskenzai, and D Litman. 2008. Comparing spoken dialog corpora collected with recruited subjects versus real users. In *Proc SIGdial, Columbus, Ohio, USA*.

AW Black, S Burger, B Langner, G Parent, and M Eskenazi. 2010. Spoken dialog challenge 2010. In *Proc SLT, Berkeley, CA*.

J Henderson and O Lemon. 2008. Mixture model POMDPs for efficient handling of uncertainty in dialogue management. In *Proc ACL-HLT, Columbus, Ohio*.

G Parent and M Eskenazi. 2010. Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *Proc SLT, Berkeley, CA*.

A Raux, B Langner, D Bohus, A Black, and M Eskenazi. 2005. Let's go public! Taking a spoken dialog system to the real world. In *Proc INTERSPEECH, Lisbon*.

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

JD Williams and S Balakrishnan. 2009. Estimating probability of correctness for ASR N-best lists. In *Proc SIGdial, London, UK*.

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, I Arizmendi, and A Conkie. 2010. Demonstration of AT&T "Let's Go": A production-grade statistical spoken dialog system. In *Proc SLT, Berkeley, CA*.

JD Williams. 2007. Using particle filters to track dialogue state. In *Proc IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Kyoto, Japan*.

JD Williams, 2010a. *AT&T Statistical Dialog Toolkit*. `http://www.research.att.com/people/Williams_Jason_D`.

JD Williams. 2010b. Incremental partition recombination for efficient tracking of multiple dialog states. In *Proc Intl Conf on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, USA*.

SJ Young, M Gašić, S Keizer, F Mairesse, J Schatzmann, B Thomson, and K Yu. 2010. The hidden information state model: a practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174, April.

Appendix: Mechanism illustrations

This appendix provides graphical illustrations of each of the four *mechanisms* that can cause the top ASR hypothesis to be different from the top belief state hypothesis. These examples were taken from logs of calls with real users, although some surface forms have been simplified for space.

At the top of each panel is the system action taken. The user's true response is shown in italics in the left-most column. The second column shows the top 7 entries from the ASR N-best list, displayed in the order produced by the speech recognition engine. The third column shows the confidence score – the local probability of correctness assigned to each ASR N-best entry. The last column shows the resulting belief state, sorted by the magnitude of the belief. Correct entries are shown in bold red.

ASR re-ranking and prior re-ranking occur within one turn, and confidence aggregation and N-best synthesis occur across two turns. These examples all show cases where the belief state is correct and the ASR is incorrect; however, the opposite also occurs of course.

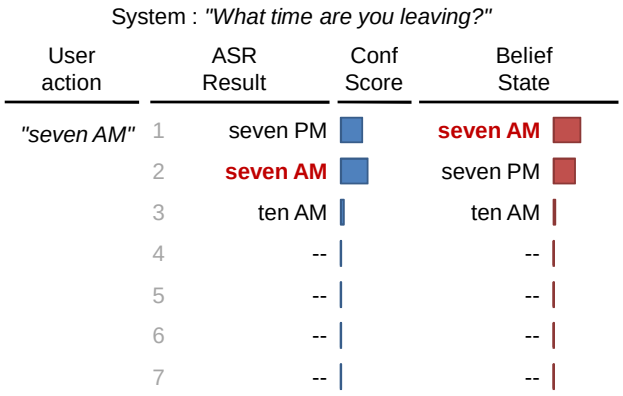

Figure 7: **Illustration of ASR re-ranking**: The correct ASR hypothesis ("seven AM") is in the $n = 2$ position, but it is assigned a higher confidence score than the misrecognized $n = 1$ entry "seven PM". TIME uses a flat prior, so the higher confidence score results in "seven AM" attaining the highest belief.

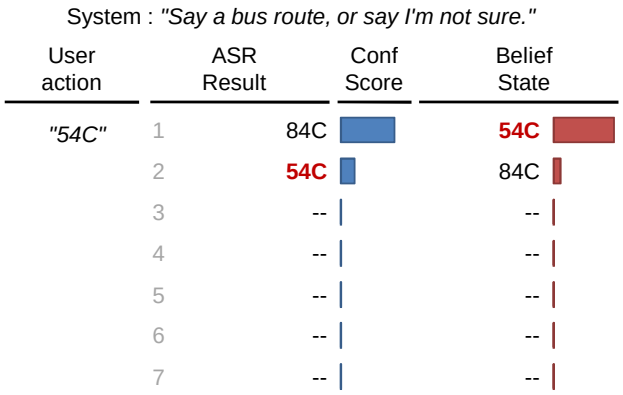

Figure 8: **Illustration of Prior re-ranking**: The correct ASR hypothesis ("54C") is in the $n = 2$ position, and it is assigned less confidence than the mis-recognized $n = 1$ entry, "84C". However, the prior on 54C is much higher than on 84C, so 54C obtains the highest belief.

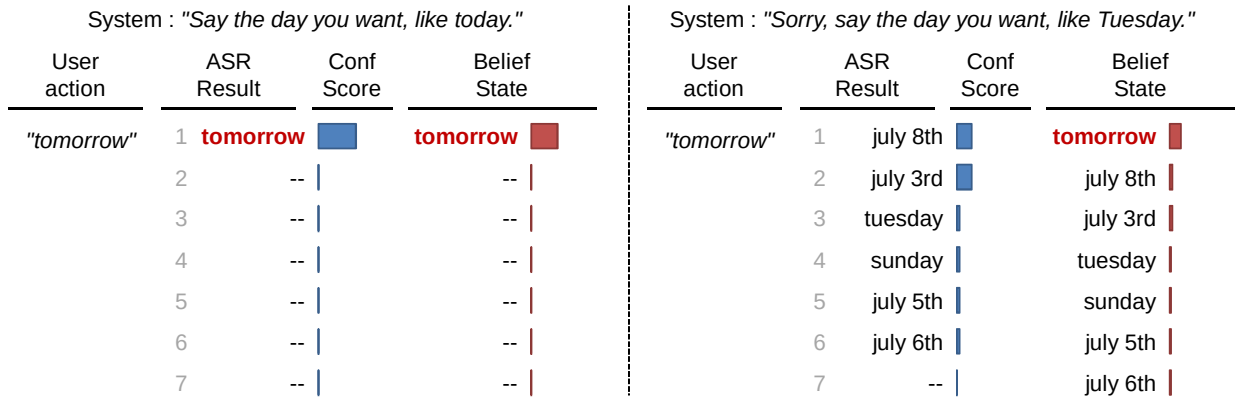

Figure 9: **Illustration of Confidence aggregation**: In the first turn, "tomorrow" is recognized with medium confidence. In the second turn, "tomorrow" does not appear on the N-best list; however the recognition result has very low confidence, so this misrecognition is unable to dislodge "tomorrow" from the top belief position. At the end of the second update, the belief state's top hypothesis of "tomorrow" is correct even though it didn't appear on the second N-best list.

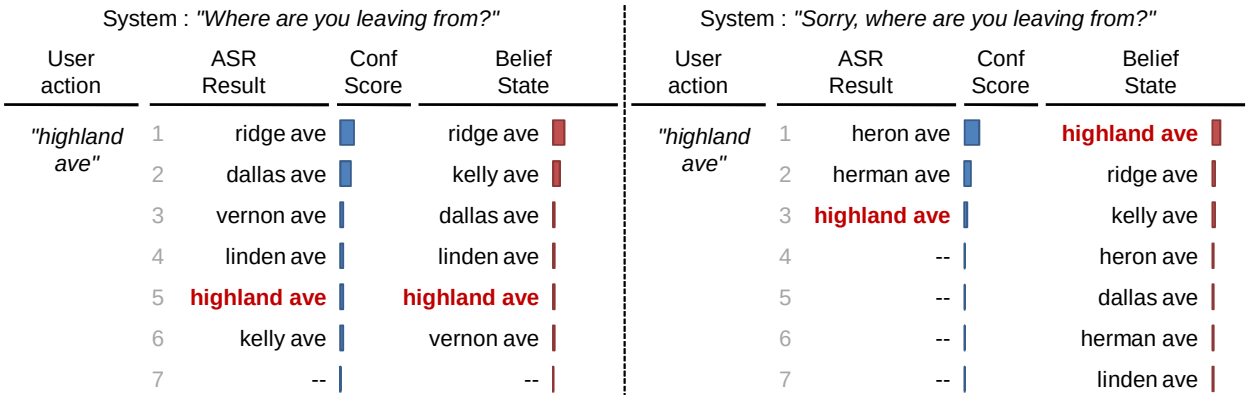

Figure 10: **Illustration of N-best synthesis**: In the first turn, the correct item "highland ave" is on the ASR N-best list but not in the top position. It appears in the belief state but not in the top position. In the second turn, the correct item "highland ave" is again on the ASR N-best list but again not in the top position. However, because it appeared in the previous belief state, it obtains the highest belief after the second update. Even though "highland ave" was mis-recognized twice in a row, the commonality across the two N-best lists causes it to have the highest belief after the second update.