Using Performance Trajectories to Analyze the Immediate Impact of User State Misclassification in an Adaptive Spoken Dialogue System

Kate Forbes-Riley
Learning Research & Development Ctr (LRDC)
University of Pittsburgh
Pittsburgh, PA 15260
forbesk@cs.pitt.edu

Diane Litman
Dept. Computer Science & LRDC
University of Pittsburgh
Pittsburgh, PA 15260
litman@cs.pitt.edu

Abstract

We present a method of evaluating the *immediate* performance impact of user state misclassifications in spoken dialogue systems. We illustrate the method with a tutoring system that adapts to student uncertainty over and above correctness. First we define a ranking of user states representing local performance. Second, we compare user state trajectories when the first state is accurately classified versus misclassified. Trajectories are quantified using a previously proposed metric representing the likelihood of transitioning from one user state to another. Comparison of the two sets of trajectories shows whether user state misclassifications change the likelihood of subsequent higher or lower ranked states, relative to accurate classification. Our tutoring system results illustrate the case where user state misclassification increases the likelihood of negative performance trajectories as compared to accurate classification.

1 Introduction

Spoken dialogue systems research has shown that natural language processing errors can negatively impact global system performance. For example, automatic speech recognition errors have been shown to negatively correlate with user satisfaction surveys taken after the system interaction is over (e.g., (Walker et al., 2000a; Pon-Barry et al., 2004)).

Automatic user state classification errors have also been shown to negatively impact global performance in spoken dialogue systems (e.g., (Pon-Barry et al., 2006)). For example, in our prior work with an uncertainty-adaptive spoken dialogue computer tutoring system, we found that recognizing and adapting to the user's state of uncertainty, over and above his/her state of correctness, significantly improved global learning over all users (as measured by tests taken before and after the system interaction). However, this was only true when the user uncertainty was manually labeled during the interaction by an unseen human "wizard of oz" (Forbes-Riley and Litman, 2011b); it was not true when the uncertainty was automatically labeled by the system. Further analysis showed that uncertainty classification errors largely accounted for the global performance decrease in our fully automated system. In particular, only a small proportion of users' actual uncertainty was being accurately classified by the system (Forbes-Riley and Litman, 2011a).[1]

The question we address in this study is how to analyze the impact of automatic user state classification errors when analyzing performance at a *local* level. In particular, is there a measurable local performance difference when one compares what happens in a dialogue after a turn is accurately classified versus misclassified? We show here how user state trajectories can be used to answer this question. First, a ranking of user states is defined (Section 3.1). Second, user state trajectories are computed from two sets of system dialogue: one in

[1] In natural language processing (NLP) research, the terms "(in)correct" and "(un)certain" can have multiple interpretations. To avoid confusion, we reserve these terms in this paper *only* to refer to the semantic content and affective/attitudinal expression of user answers (respectively). When referring to the NLP performance of our system, we use the terms "accurately classified" and "misclassified".

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 216–226,
Portland, Oregon, June 17-18, 2011. ©2011 Association for Computational Linguistics

which the user state of interest is accurately classified in the first turn in the trajectory, and another in which it is misclassified (Section 3.2). Trajectories are quantified as the likelihood of transitioning from one user state to another (D'Mello et al., 2007). Comparison of the two sets of trajectories indicates how user state misclassifications change the relative likelihood of subsequent states. Transitions to higher ranked states indicate improved local performance while transitions to lower ranked states indicate decreased local performance.

In our research, we are interested in this question because we hypothesize that accurate and inaccurate user state classification in our uncertainty-adaptive system yielded immediate differences in user behavior. We further hypothesize that our uncertainty-adaptive system had a negative immediate impact on the user's state when (un)certainty was misclassified, as compared to when (un)certainty was accurately classified. Our user state trajectory results support these hypotheses. We find that (un)certainty misclassifications increased the likelihood of transitioning to the lowest ranked user state in the next turn. In contrast, accurate (un)certainty classification yielded an increased likelihood of more positive performance trajectories (Section 4).

More generally, this question is relevant to other automatically classified user states and other types of dialogue systems, whenever the goal is to understand the immediate impact of user state classification errors on user behavior during the dialogue (Sections 3.1 and Section 5).

2 The System and Dialogues

We apply this local performance analysis to dialogues between college students and our fully automated spoken dialogue tutoring system, ITSPOKE.[2]

Two sets of dialogues are used here, which come from two versions of ITSPOKE: the uncertainty-adaptive and non-adaptive versions. Both versions automatically classify user (un)certainty and (in)correctness for each turn. However, the non-adaptive version's responses are based only on (in)correctness, while the uncertainty-adaptive version provides an uncertainty adaptation to uncertain+correct answers. All dialogues were collected in our prior experiment comparing global learning across the uncertainty-adaptive and non-adaptive system versions (see Section 1). The uncertainty-adaptive system yielded 120 dialogues (1957 student turns) from 24 subjects. The non-adaptive system yielded 125 dialogues (2065 student turns) from 25 subjects. Our analysis will focus on the dialogues from the uncertainty-adaptive system (Section 4.1); the dialogues from the non-adaptive system will be used for comparison (Section 4.2).

Briefly, ITSPOKE tutors 5 physics problems (one per dialogue), in a Tutor Question - Student Answer - Tutor Response format. The tutor questions serially present topics needed to solve each problem; adjacent questions concern identical or closely related topics. After each tutor question, the student answer is digitized from head-mounted microphone input and sent to the Sphinx2 recognizer, which yields an automatic transcript. The answer's (in)correctness is then automatically classified based on this transcript, using the TuTalk semantic analyzer (Jordan et al., 2007). Simultaneously, the answer's (un)certainty is automatically classified by inputting features of the speech signal, the automatic transcript, and the dialogue context into a logistic regression model.[3] All natural language processing components were trained using prior ITSPOKE corpora.

The dialogue manager determines the appropriate tutor response and sends it to the Cepstral text-to-speech system. The uncertainty-adaptive and non-adaptive system versions give the same response to incorrect answers (regardless of (un)certainty): incorrectness feedback (e.g., "That's not right") followed by additional remedial content to help students understand why their answer was incorrect. Both versions also give the same response to certain+correct answers: correctness feedback only (e.g., "That's right"). The two versions differ in their response to uncertain+correct answers. The non-adaptive version responds with correctness feedback only, while the uncertainty-adaptive version responds with correctness feedback followed by additional remedial content to help students resolve their uncertainty. This uncertainty adaptation is based on

[2] ITSPOKE(**I**ntelligent **T**utoring **SPOKE**n dialogue) is a spoken, modified version of Why2-Atlas (VanLehn et al., 2002).

[3] The uncertainty model includes lexical, pitch, temporal, and energy features as well as tutor question and user gender, and is described further in (Forbes-Riley and Litman, 2011a).

tutor1: Let's begin by looking at the motion of the man and his keys while he's holding them. How does his velocity compare to that of his keys?

student-automatic1: his also the is the same as that of his keys points force **(certain+incorrect)**

student-manual1: his velocity is the same as that of his keys while he's holding them (uncertain+correct)

tutor2: Actually... Their velocities must be the same, because the man's holding his keys, so they're not moving with respect to each other. [...] What are the forces exerted on the man after he releases his keys?

student-automatic2: gravity **(uncertain+correct)**

student-manual2: gravity?? (uncertain+correct)

non-adaptive-tutor3: Right. So what's the direction of the force of gravity on the man (e.g., vertically up)?

uncertainty-adaptive-tutor3: Right. Gravity is the only force acting on the man. This force is also called the weight of the man. It's always present when an object is in earth's gravitational field. So what's the direction of the force of gravity on the man (e.g., vertically up)?

Figure 1: Annotated Dialogue Example Showing Uncertainty Adaptation and User State Classifications

prior tutoring research, which argues that incorrectness and uncertainty are both signals of a "learning impasse", i.e., an opportunity to better learn the material (VanLehn et al., 2003). Our uncertainty adaptation only provides additional remedial content after the uncertain+correct impasse, because incorrectness impasses (uncertain and certain) already receive remedial content from the non-adaptive system. It is assumed that this content helps resolve both incorrectness and uncertainty (when present).

After the experiment, each student answer (turn) was manually transcribed and labeled for (un)certainty and (in)correctness. One labeler performed the annotation based on schemes developed and evaluated on prior ITSPOKE corpora, where this labeler and another labeler displayed interannotator reliability of 0.85 and 0.62 Kappa on (in)correctness and (un)certainty, respectively (Forbes-Riley and Litman, 2011a).[4] Comparison of the automatic and manual labels yielded 84.7% accuracy for automatic (in)correctness classification and 80.3% accuracy for automatic (un)certainty classification. However, the (un)certainty model had an uncertainty recall of only about 20%, while the (in)correctness model had a correctness recall of about 80% (Forbes-Riley and Litman, 2011a).[5]

Figure 1 illustrates ITSPOKE's natural language processing components and the two system versions. The first answer is classified as certain+incorrect (**student-automatic1**) but manually labeled as uncertain+correct (*student-manual1*); the manual and automatic transcripts are also substantially different. Because this answer was misclassified as incorrect, both versions give the same response (**tutor2**). The second answer is accurately classified as uncertain+correct. The non-adaptive system thus ignores the uncertainty and only provides correctness feedback (**non-adaptive-tutor3**), while the adaptive system responds with correctness feedback and additional remedial content to help resolve the uncertainty (**uncertainty-adaptive-tutor3**).

3 Local Performance Evaluation

Here we discuss how to evaluate the local impact of user state misclassification in dialogue systems.

3.1 Defining a User State Severity Ranking

Building on tutoring research that views both uncertainty and incorrectness as signals of learning impasses (Section 2), we previously defined a severity ranking for the four impasse states corresponding to all combinations of binary (in)correctness

[4] Because these evaluations showed that this trained labeler could reliably annotate (un)certainty and (in)correctness in ITSPOKE dialogues, no further evaluations were performed.

[5] The lower recall for predicting uncertainty is nevertheless higher than always predicting no uncertainty (a majority class baseline has 0% recall), and is on par with prior work in affect-adaptive tutoring systems, e.g. (Walonoski and Heffernan, 2006); in general affective systems research has found it difficult to accurately predict positive occurrences of affect.

Impasse State:	**certain+incorrect**	**uncertain+incorrect**	**uncertain+correct**	**certain+correct**
Severity:	*most*	*less*	*least*	*none*

Figure 2: User Impasse State Severity Ranking

and (un)certainty (Forbes-Riley and Litman, 2011a). This ranking, shown in Figure 2, reflects the assumption that a student must perceive an impasse in order to resolve it. A state of uncertainty reflects this awareness. Therefore, the most severe type of learning impasse occurs when a student is incorrect but not aware of it. Impasse states of decreasing severity occur when the student is incorrect but aware that s/he might be, and correct but believes s/he may not be, respectively. No impasse exists when a student is correct and not uncertain about it.

In our prior work, this ranking of user states was independently validated by showing that average impasse state severity negatively correlates with global learning gain in our system dialogues (Forbes-Riley and Litman, 2011a). In other words, a higher proportion of user states with less severe or no impasses directly relates to higher global learning gain.

More generally, the idea of ranking user states in terms of those that do or do not represent *communication impasses* applies to other dialogue system domains and other user state dimensions as well. For example, in information-seeking domains, frustration and anger are common affective states whose occurrence during the dialogue signals severe communication problems (Batliner et al., 2003), while hang-ups and turns requesting a human operator are other types of user states whose occurrence during the dialogue signals severe communication problems (Walker et al., 2000b).

Moreover, state trajectories can be used to represent abstractions over other types of user (or system) behaviors. In our tutoring system analysis, representing user states in terms of only (un)certainty and (in)correctness is an abstraction that we find useful for analyzing impasse trajectories. However, during run-time, a finite-state dialogue manager consisting of 142 states actually controls the system's operation, and uses many other features besides user uncertainty and incorrectness to determine the system's response (e.g. the physics concepts related to the current system question, the history of prior student answers to similar questions, etc.). Any of these states could be analyzed as well to understand their local performance impact, as could their analogs in other system domains. For example, in a train dialogue system, while the actual state representation used during operation could be quite complex, for a trajectory analysis a simpler representation could be suitable, one which tracks whether the system knows the values of the n attributes needed to query the database. The state ranking in this case would be over equivalence classes of states: states with n attributes known > states with n-1 attributes known > ... > initial state with 0 attributes known.

3.2 Computing User State Trajectories

Local trajectories of user states during a dialogue can be computed as the likelihood of transitioning from the user state in turn *n* to the user state in turn *n+1*. Here we use D'Mello et al.'s metric, *transition likelihood L* (D'Mello et al., 2007).

Transition likelihood L is computed as shown below, where *n* refers to the impasse state in turn *n* and *n+1* refers to the impasse state in turn *n+1*. As shown, L is computed as the conditional probability that the user state in turn *n+1* will occur given that the user state in turn *n* has occurred, adjusted for the base rate of occurrence of the user state in turn *n+1*. The denominator normalizes the result so that L ranges from $-\infty$ to 1. L=1 indicates that *n+1* always follows *n* over and above the probability of n+1 occurring. L=0 indicates that *n+1* follows *n* at the chance level. L<0 indicate that the likelihood of *n+1* following *n* is much lower than the base rate of *n+1* occurring.[6]

$$\mathbf{L(n \rightarrow n+1)} = \frac{P(n+1|n)-P(n+1)}{1-P(n+1)}$$

Transition likelihood L has previously been used to compute the likelihood of transitioning from one affective state to another (e.g., from confusion to

[6] Note that this metric, which assesses the adjusted probability of one user state following another, is equivalent to Kappa in computing agreement among annotators after adjusting for chance (D'Mello et al., 2007).

frustration) in a single set of dialogues between student and computer tutor (D'Mello et al., 2007). Transition likelihood L has also been used to compare how the likelihoods of transitioning from one affective state to another vary across two different sets of dialogues collected with two different versions of an affect-adaptive tutoring system (McQuiggan et al., 2008). Our analysis is based on this analysis, but extends it in three ways: 1) our transitions involve complex user states composed of two dimensions ((un)certainty and (in)correctness), 2) the user states in our transitions are ranked to enable a local performance analysis, 3) our performance analysis is applied to the question of how user state misclassification impacts local performance, by comparing transition likelihoods after accurate and inaccurate user state classifications.

In this prior work and in our work, likelihoods for each transition are computed for each user (over all dialogues of a user). ANOVAs with post-hoc pairwise tests can then determine if there were significant differences between all possible transitions from the current user state in turn *n*.

To investigate how user state misclassifications impact local performance, *two* user trajectories are computed per user for each *n*→*n+1* transition: one when the manual and automatic user state labels for turn *n* agreed, and another when they did not agree. In both cases, using the manual label for turn *n+1* enables the *true* final user state to be compared across the two sets of trajectories. Comparison of the final state in the two sets of trajectories indicates how user state misclassifications change the relative likelihood of the subsequent user states. Transitions to higher ranked states indicate improved local performance while transitions to lower ranked states indicate decreased local performance.

4 Impact of User State Misclassifications in Uncertainty-Adaptive ITSPOKE

We now apply this analysis to the uncertainty-adaptive ITSPOKE dialogues, to investigate how user state misclassification impacts the local performance of the uncertainty adaptation.

Since the complex user state of uncertain+correct triggers the uncertainty adaptation, misclassifying (un)certainty or (in)correctness can potentially impact the local performance of the adaptation. However, as noted in Section 2, we previously found that uncertainty misclassifications in our system were more severe than correctness misclassifications. Thus, to streamline our analysis and avoid data skew issues, we focus on how (un)certainty misclassifications in manually labeled correct answers impact our local performance trajectories. There are 1270 manually labeled correct turns in the dialogues collected with uncertainty-adaptive ITSPOKE. In the dialogues collected with non-adaptive ITSPOKE (which we will use for comparison), there are 1353 manually labeled correct turns.

We hypothesize that when (un)certainty misclassification in correct answers causes the uncertainty adaptation to be erroneously triggered or blocked, we will see a negative performance impact, in terms of an increased likelihood of transitioning to a more severe impasse state when uncertainty is misclassified as compared to when it is accurately classified.

4.1 Uncertainty-Adaptive ITSPOKE Results

Accurate Uncertainty Classification: Figure 3 presents descriptive statistics for the likelihood (L) that a manually labeled uncertain+correct answer *accurately classified as uncertain* in turn *n* will transition to each of the four manually labeled impasse states in turn *n+1*. As noted in Section 3.2, L=0 indicates that the transition likelihood is equal to chance, while L>0 and L<0 indicate likelihoods greater and less than chance, respectively.

An ANOVA indicated that there were statistically significant differences among the likelihoods in Figure 3 (F(3,56)=3.87, p=.02). The most likely transitions are shown with stripes. Specifically, post-hoc pairwise tests showed that in turn *n+1*, an uncertain+incorrect answer (p<.01) or uncertain+correct answer (p=.02) is significantly more likely than a certain+correct answer (but are themselves equally likely). In addition, an uncertain+incorrect answer is significantly more likely than a certain+incorrect answer (p=.05), in turn *n+1*. A dialogue example of the most likely transition after accurately classified uncertainty is shown in Figure 5, where it is compared with the misclassified minimal pair in Figure 6 (see Appendix).

These results indicate that accurately classifying (and thus accurately adapting to) uncertain+correct

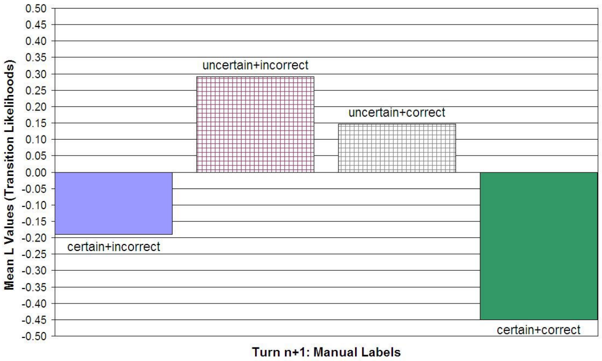

Figure 3: Turn *n* → Turn *n+1* Transition Likelihoods (L) after a manually labeled uncertain+correct answer in turn *n* is accurately classified as uncertain and receives the uncertainty adaptation

answers is most likely to yield continued uncertainty (regardless of correctness) in turn *n+1*. Prior research (Craig et al., 2004; Kort et al., 2001) has shown that uncertainty and questioning are positive and crucial aspects of the learning process. The continued uncertainty suggests that the uncertainty adaptation keeps the student engaged in the learning process, and the equal likelihood of correctness or incorrectness accompanying this uncertainty suggests that they have not yet unreservedly adopted either the correct or incorrect line of reasoning about the topic under discussion.

To determine whether any of these transitions are directly tied to global performance, we computed Pearson's correlations over all students between the percentage of each transition and global learning gain.[7] Interestingly, transitioning from an accurately classified correct+uncertain answer to a correct+certain answer is negatively related to global learning gain (R=-.458, p=.025). This indicates that continued uncertainty after the uncertainty adaptation is provided is more beneficial, in the long run, than no uncertainty. No other trajectories are directly related to global learning. Although our prior result, that average impasse severity negatively correlates with global learning gain (Section 3.1), indicates it is better from a global perspective for a student to *be* in a state of no impasse (correct+certain), it does not tell us the best way for the student to attain this state. The results of our transition correlations shed light on this - they tell us that transitioning directly from correct+uncertain is not the best way to attain the no impasse state. We hypothesize that looking at wider transition windows (e.g., trigrams) will shed light on what *is* the best way to attain this state. For example, it may be that the best way to transition to a state of no impasse is to do so after sustained uncertainty (as in Figure 3).

Uncertainty Misclassification: Considering now user state misclassifications, our results for accurately classified uncertain+correct answers are in sharp contrast to those for manually labeled uncertain+correct answers *misclassified as certain* in turn *n*. In particular, an ANOVA indicated that all manually labeled impasse states are equally likely in *n+1* (F(3,88)=1.22, p=.32) after a misclassified uncertain+correct answer.[8]

These results indicate that misclassifying (and erroneously *not* adapting to) uncertain+correct answers is as likely to have an immediate negative impact on learning as it is to have a neutral or positive impact. In particular, the misclassification is likely to cause some students to transition from the least severe impasse about the concept in turn *n* to the most severe impasse about the concept in turn *n+1*.[9] When they do not receive the uncertainty adaptation, these students adopt an incorrect line of reasoning in turn *n+1*, without any uncertainty about it at all.

As illustration, compare the example in Figure 5, where uncertainty is accurately classified, with the example in Figure 6, where uncertainty is misclassified (see Appendix). As shown, the uncertainty in *student-manual1* signals that further explanation is needed. When received (Figure 5) the student still makes a math error on the next question, but s/he appears to understand the task. In contrast, when the uncertainty adaptation is erroneously not received (Figure 6), there is no indication that the student's understanding has increased; s/he appears to be simply repeating the number 9.8 (a number which appears frequently in Newtonian physics). User uncertainty misclassification in other domains could have

[7] normalized learning gain = (posttest-pretest)/(1-pretest).

[8] Since the ANOVA results were non-significant, no figure or correlations are discussed.

[9] As noted in Section 2, adjacent turns within a dialogue will either address the same or closely related topics.

similar effects; in general, if a user is uncertain in turn *n* about how to perform a task, and the system moves on without supplying information to resolve this uncertainty, there may be an immediate negative impact if that knowledge is required or presupposed again in turn *n+1*.

Accurate Certainty Classification: Turning now to manually labeled certain+correct answers, Figure 4 presents descriptive statistics for the likelihood that when *accurately classified as certain* in turn *n*, certain+correct answers will transition to each of the four manually labeled impasse states in turn *n+1*. An ANOVA indicated that there were statistically significant differences among these likelihoods (F(3,92)=17.96, p<.01). The most likely transitions are shown with stripes. More specifically, post-hoc pairwise tests showed that in turn *n+1*, a manually labeled certain+correct answer is significantly more likely than any other impasse state (p<.01), and all other impasse states were equally likely. A dialogue example of the most likely transition after accurately classified certainty is shown in Figure 7, where it is compared with the misclassified minimal pair in Figure 8 (see Appendix).

These results indicate that accurately classifying and *not* adapting to certain+correct answers has an immediate positive impact on the learning process, by not introducing learning impasses about concepts already understood. Note however that Pearson's correlations for these transitions showed no significant relation to global performance.

Certainty Misclassification: Again, our results for accurately classified certain+correct answers are in sharp contrast with those found for manually labeled certain+correct answers *misclassified as uncertain* in turn *n*. An ANOVA indicated that all manually labeled impasse states are equally likely in turn *n+1* (F(3,72)=0.33, p=.80). These results indicate that misclassifying and erroneously *adapting* to certain+correct answers is as likely to have an immediate negative impact on learning as it is to have a neutral or positive impact. In particular, the misclassification is likely to cause some students to transition from no impasse to the most severe impasse state. When they erroneously receive the uncertainty adaptation, these students go from no impasse at all in turn *n* to an incorrect line of reasoning in turn *n+1*, without any uncertainty about it at all.

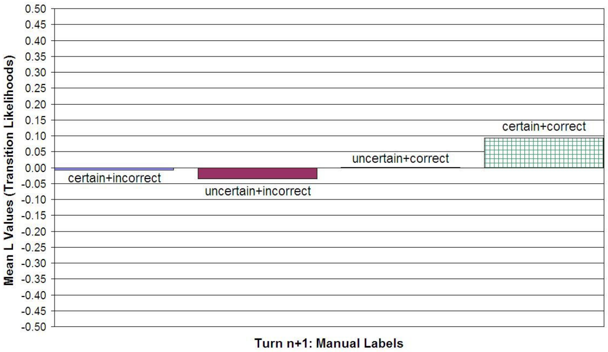

Figure 4: Turn *n* → Turn *n+1* Transition Likelihoods (L) after a manually labeled certain+correct answer in turn *n* is accurately classified as certain and does not receive the uncertainty adaptation

As illustration, compare the example in Figure 7, where certainty is accurately classified, with the example in in Figure 8, where certainty is misclassified (see Appendix). As shown, the certainty in *student-manual1* signals that no further explanation is needed so the system can move on (Figure 7). When the uncertainty adaptation is erroneously received even though the student is certain (Figure 8), this appears to have caused the student to stop paying close attention and thus provide an obviously incorrect answer to an easy question. User certainty misclassification in other domains could have similar effects; in general, if a user is already certain in turn *n* about how to perform a task, and the system "wastes" his/her time by resupplying information that is already understood, there may be an immediate negative impact in terms of loss of focus, disengagement, or even decreased understanding, that cause the task in turn *n+1* to be performed incorrectly.

4.2 Comparing *Non-Adaptive* ITSPOKE

As a sanity check, we performed the same trajectory analysis on the dialogues from the *non-adaptive* version of the system. The purpose here was to confirm the presupposition of the above analysis, that uncertainty-adaptive ITSPOKE was actually producing different local behaviors than non-adaptive ITSPOKE. In other words, since the non-adaptive system ignores uncertainty, there should be no difference in transition likelihoods when uncertainty is accurately classified versus when it is misclassified.

This expectation was borne out. ANOVAs indicated that in the non-adaptive system, a manually labeled uncertain+correct answer is equally likely to transition to any of the four manually labeled impasse states in turn *n+1*, regardless of whether it was accurately classified as uncertain in turn *n* (F(3,48)=0.25, p=.86) or misclassified as certain in turn *n* (F(3,92)=0.07, p=.98). Thus as expected, uncertain+correct answers in the non-adaptive system pattern like uncertain+correct answers *misclassified as certain* in the uncertainty-adaptive system. In both cases, we see the same negative immediate performance impact of not giving uncertain+correct answers the uncertainty adaptation.

ANOVAs with post-hoc pairwise tests further indicated that in the non-adaptive system, a manually labeled certain+correct answer is significantly more likely to transition to a certain+correct answer than to any other manually labeled impasse state, regardless of whether it was accurately classified as certain in turn *n* (ANOVA:(F(3,96)=20.81, p<.001), post-hoc tests: p<.001) or misclassified as uncertain in turn *n* (ANOVA:(F(3,80)=14.00, p<.001), post-hoc tests: p<.001). Thus as expected, certain+correct answers in the non-adaptive system pattern like *accurately classified* certain+correct answers in the uncertainty-adaptive system. In both cases, we see the same positive immediate performance impact of not giving manually labeled certain+correct answers the uncertainty adaptation.

4.3 Comparing Local and Global Performance Results

Finally, in analyses such as this one, comparing local and global performance results can help pinpoint specific areas for future system redesign. In our case, this comparison suggests the most important aspect to focus on with respect to improving our uncertainty model.

In particular, as noted in Section 1, we previously found that the low uncertainty recall of our system (approximately 20%) had a negative global performance impact; mistaking so much true uncertainty for certainty substantially reduced the amount users learned (Forbes-Riley and Litman, 2011a). We also showed in this prior work that mistaking certainty for uncertainty did not negatively impact the amount users learned. These results suggested that the system should be less cautious in applying the uncertainty-adaptive behavior; i.e., applying it whenever there is some possibility that the user is actually uncertain, even if it means applying it to some turns that are actually certain.

On the other hand, our local performance analysis in this paper showed that (un)certainty misclassification increased the likelihood of an immediate negative impact on learning. These results suggest that the system should be more cautious in applying the uncertainty-adaptive behavior; i.e., only applying it when there is a high probability that the user is actually uncertain.

Together these local and global results suggest that we should focus on improving uncertainty recall without decreasing uncertainty precision, in our uncertainty model. With this goal in mind, we are currently exploring the use of features and methods from recent INTERSPEECH emotion and paralinguistic challenges (Schuller et al., 2009; Schuller et al., 2010).

5 Conclusion and Future Directions

This paper presents an approach for analyzing the immediate impact of user state misclassifications in dialogue systems. A ranking of user states is defined, and then user state trajectories are compared when the first state is accurately classified versus misclassified. Trajectories are quantified using a previously proposed metric representing the likelihood of transitioning between states. Comparison of the two sets of trajectories shows whether misclassifications change the likelihood of subsequent higher or lower ranked states, relative to accurate classification. We illustrated the approach with an adaptive tutoring system that automatically detects and adapts to student uncertainty.

As our results indicate, the approach can be used to answer questions which global performance analyses overlook. First, the analysis shows whether user state misclassifications actually matter locally - whether these errors have an immediate effect on user behavior or not. Moreover, the analysis can determine whether this effect is positive or negative or neutral. In our tutoring system data, we found that misclassifying user uncertainty had a negative immediate impact on user behavior, relative to accurate classification.

The analysis can also confirm that a dialogue intervention actually changes user behaviors. In our tutoring system data, we found that the adaptive system yielded significantly different user state trajectories than the non-adaptive system, even though, as noted in Section 1, our prior global performance analysis did not show any overall differences among the global performance metrics that we examined across the adaptive and non-adaptive systems.

In addition, the analysis can confirm that a dialogue intervention shifts user behaviors in the desired direction. In our tutoring system data, we found that the immediate effect of accurately adapting to uncertainty was most likely to be continued uncertainty. Although the adaptation does not yield an immediate transition to the highest ranked user state, the outcome is clearly more positive than that of ignoring uncertainty, which increases the likelihood of transitioning to the lowest ranked user state.

Finally, the local performance results can shed light on the steps needed to improve global performance, by investigating how the two are related. In our tutoring system data, we found that there is not a one-to-one relationship between the most beneficial local and global outcomes. In particular, transitioning directly to the highest ranked (no impasse) state after receiving the uncertainty adaptation was negatively correlated to global learning gain. We hypothesized that looking at wider transition windows (e.g., trigrams) will shed light on what *is* the best local path to the highest ranked state.

We conclude by emphasizing that state trajectories can be used to represent abstractions over various types of user (or system) behaviors, in various domains, whenever their local performance impact is viewed as important to understand.

Acknowledgments

This work is funded by NSF awards #0914615 and #0631930. We thank Reva Freedman and the IT-SPOKE group for comments.

References

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth. 2003. How to find trouble in communication. *Speech Communication*, 40:117–143.

S. Craig, A. Graesser, J. Sullins, and B. Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29(3).

S. D'Mello, R. S. Taylor, and A. Graesser. 2007. Monitoring affective trajectories during complex learning. In *Proc. Cognitive Science Society*.

K. Forbes-Riley and D. Litman. 2011a. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*. In Press.

K. Forbes-Riley and D. Litman. 2011b. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech and Language*, 25(1):105–126.

P. Jordan, B. Hall, M. Ringenberg, Y. Cui, and C.P. Rose. 2007. Tools for authoring a dialogue agent that participates in learning studies. In *Proc. Artificial Intelligence in Education*.

B. Kort, R. Reilly, and R. Picard. 2001. An affective model of interplay between emotions and learning : Reengineering educational pedagogy-building a learning companion. In *Proc. IEEE Conference on Advanced Learning Technology*.

S. W. McQuiggan, J. L Robison, and J. C. Lester. 2008. Affective transitions in narrative-centered learning environments. In *Proc. Intelligent Tutoring Systems Conference*.

Heather Pon-Barry, Brady Clark, Elizabeth Owen Bratt, Karl Schultz, and Stanley Peters. 2004. Evaluating the effectiveness of SCoT:a spoken conversational tutor. In *Proceedings of ITS Workshop on Dialogue-based Intelligent Tutoring Systems*.

H. Pon-Barry, K. Schultz, E. Bratt, B. Clark, and S. Peters. 2006. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education*, 16.

B. Schuller, S. Steidl, and A. Batliner. 2009. The Interspeech 2009 emotion challenge. In *Proc. Interspeech*.

B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan. 2010. The INTERSPEECH 2010 paralinguistic challenge. In *Proc. Interspeech*.

K. VanLehn, P. Jordan, C. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, R. Srivastava, and R. Wilson. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. Intelligent Tutoring Systems Conference*.

K. VanLehn, S. Siler, and C. Murray. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3).

Marilyn A. Walker, Candace A. Kamm, and Diane J. Litman. 2000a. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3).

Marilyn A. Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman. 2000b. Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You? In *Proceedings of the 1st Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL)*.

J. A. Walonoski and N. T. Heffernan. 2006. Detection and analysis of off-task gaming behavior in intelligent tutoring systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS)*, pages 382–391, Jhongli, Taiwan.

APPENDIX

tutor1: If an object in freefall begins with a velocity of zero, what'll its velocity be a second later? Please specify the units too.

student-automatic1: 9.8 meters per second **(uncertain+correct)**

student-manual1: 9.8 meters per second?? (uncertain+correct)

tutor2: Fine. Because it's in freefall, its acceleration is 9.8m/s^2. This tells us that every second, its velocity increases by 9.8 m/s. Therefore, at one second its velocity is 0 m/s, and a second later, its velocity is 9.8m/s. So, what'll its velocity be a second after that?

student-manual2: 18.6 meters per second?? (uncertain+incorrect)

Figure 5: Annotated Dialogue Example of Most Likely Transition after Accurate Uncertainty Classification

tutor1: If an object in freefall begins with a velocity of zero, what'll its velocity be a second later? Please specify the units too.

student-automatic1: 9.8 meters per second **(certain+correct)**

student-manual1: 9.8 meters per second?? (uncertain+correct)

tutor2: Fine. So, what'll its velocity be a second after that?

student-manual2: 9.8 meters per second (certain+incorrect)

Figure 6: Annotated Dialogue Example of Negative Impact of Uncertainty Misclassification

tutor1: There's only one horizontal force exerted on the truck during the collision with the car. What is it?

student-automatic1: impact force **(certain+correct)**

student-manual1: the impact force (certain+correct)

tutor2: Right. Now let's find the forces exerted on the truck in the vertical direction during the collision. First, what vertical force is always exerted on an object near the surface of the earth?

student-manual2: gravity (certain+correct)

Figure 7: Annotated Dialogue Example Showing Most Likely Transition after Accurate Certainty Classification

tutor1: There's only one horizontal force exerted on the truck during the collision with the car. What is it?

student-automatic1: impact force **(uncertain+correct)**

student-manual1: the impact force (certain+correct)

tutor2: Right. The problem statement says the car and the truck are both traveling in a straight horizontal line when they collide. So the impact force of the truck on the truck must be in the horizontal direction. It also says they are driving on a patch of frictionless ice when they collide. So there is no friction to take into account. This means that the impact force is the only horizontal force on the truck. Now let's find the forces exerted on the truck in the vertical direction during the collision. First, what vertical force is always exerted on an object near the surface of the earth?

student-manual2: normal force (certain+incorrect)

Figure 8: Annotated Dialogue Example Showing Negative Impact of Certainty Misclassification