ISCA Archive
http://www.isca-speech.org/archive

ITRW on Prosody in
Speech Recognition and Understanding
Molly Pitcher Inn, Red Bank, NJ, USA
October 22–24, 2001

Detecting Misrecognitions and Corrections in Spoken Dialogue Systems from 'Aware' Sites

Julia Hirschberg, Diane Litman and Marc Swerts

AT&T Labs–Research
Florham Park, NJ, 07932 USA
julia@research.att.com

University of Pittsburgh
Pittsburgh, PA 15260 USA
litman@cs.pitt.edu

IPO, Eindhoven, The Netherlands,
and CNTS, Antwerp, Belgium
m.g.j.swerts@tue.nl

Abstract

We explore the extent to which misrecognitions and corrections in spoken dialogue systems can be predicted from information about other turn categories in the preceding or following context. Features including whether or not subsequent turns represent 'aware' sites, in which users first become aware that the system has misheard them, or corrections, and whether or not prior turns represent 'aware' sites or are misrecognized are used to identify following turns as potential correction sites and to detect whether previous turns were in fact misrecognized. This represents a new phase in our ongoing work identifying corrections and misrecognitions to improve the performance of spoken dialogue systems.

1. Introduction

This paper describes new results in our continuing investigation on the prosodic reaction of users to recognition errors in Spoken Dialogue Systems (SDS). To date, we have explored whether prosodic features of user turns can tell us a) whether a speech recognition error has occurred (e.g. System hears "I want to go to Baltimore" when a user has said "I want to go to Boston") [1, 2]; b) whether a user is reacting to evidence of such a system error (e.g. System: "Did you say you want to go to Baltimore?" User: "NO!") [3]; and c) whether a user is in fact correcting such a recognition error (e.g. User: "I want to go to BOSTON!") [4, 5]. We have already found that prosodic features do predict recognition errors directly with considerable accuracy in the TOOT train information corpus dialogues. Using machine learning techniques, we have shown that, in combination with information already available to the recognizer, such as acoustic confidence scores, grammar, and recognized string, prosodic information can distinguish speaker turns that are misrecognized far better than traditional methods for ASR rejection using acoustic confidence scores alone (8.64% vs. 18.91% estimated error). More recently, we have demonstrated that misrecognitions and corrections are both prosodically distinct from what we have termed "aware" sites, defined as turns where a user first becomes aware that the system has made an error. Machine learning experiments show that these aware sites can be distinguished from other user turns with 12.2% estimated classification error. Finally, we have examined user corrections of system errors in the TOOT corpus, and have found significant prosodic differences between corrections and non-corrections that can be used to predict that a turn represents a user's correction of a system error with some success (15.72% estimated error); in addition we have uncovered interesting correlations between system strategies and types of user corrections, as well as evidence for what types of corrections are more successful, which will be important in building more successful SDS.

In this paper, we explore whether we can gain in prediction performance by combining predictors. Given that we can predict aware sites with some accuracy, we now use these predictions as predictors of misrecognitions and of corrections. We hypothesize that aware sites can function as both backward-looking and forward-looking signaling cues, making it clear to the system that something has gone wrong in the preceding context [6] and signaling corrections to come. We will describe experiments that use aware site predictions in combination with other predictors in both these ways. We explore the usefulness of features derived from hand-labeled turn classification in predicting other turn categories, to provide an upper bound on potential performance, and also present preliminary results based on predicted turn classification.

2. The TOOT Corpus

Our corpus is the TOOT SDS, which provides access to train information over the phone [7]. Subjects were 39 students, 20 native speakers of standard American English and 19 non-native speakers; 16 subjects were female and 23 male. They were asked to perform four train information seeking tasks; the exchanges were recorded and the system and user behavior logged automatically. We examined 2328 user turns from 152 dialogues generated during these experiments.

Dialogues were manually transcribed and user turns automatically compared to the corresponding ASR (one-best) recognized string to produce a *word accuracy* score (WA) for each turn. If there were any differences between the ASR output and the transcription (WA < 1), the turn was labeled as a *WA-based misrecognition*. Each turn's *concept accuracy* (CA) was also labeled by the experimenters from the dialogue recordings and the system log. If the recognizer correctly captured all the task-related information given in the user's original input (e.g. date, time, departure or arrival cities), the turn was given a CA score of 1, indicating a semantically correct recognition. Otherwise (CA < 1), the CA score reflected the percentage of correctly recognized task concepts in the turn, and the turn was labeled as a *CA-based misrecognition*. For example, if the user said "I want to go to Baltimore on Saturday at ten o'clock" but the system's best hypothesis was "Go to Boston on Saturday", the CA score for this turn would be .33. 30% of the 2328 turns were CA-based misrecognitions, while 39% were WA-based misrecognitions.

In addition to identifying turn categories representing user reactions to such misrecognitions, two authors labeled each turn as to whether or not it constituted a *correction* of a prior system failure (and if so what turn was being corrected) or represented an *aware site* for a prior failure (and if so which turn the system had failed on). Labeler disagreement was subsequently resolved by consensus. The TOOT dialogue fragment in Figure 1 illustrates these labels. 30% of the turns in our corpus were classified in this way as 'aware' sites and 29% of turns were classified as corrections. Note that aware sites may or may not also be corrections, since a user might not immediately provide correcting information. Turns that were only corrections represent 13% of the turns in our corpus; turns that were only awares represent 14%; and turns that were both account for 16%; 57% of the turns in the corpus were neither awares nor corrections. These hand-labeled turn classes as well as predicted versions of them, together with information about which turns were misrecognized by the TOOT ASR system form the classes examined below.

3. Machine Learning Experiments

For our machine learning experiments, we use RIPPER [8], a machine learning program which takes as input the classes to be learned, a set of features, and training data specifying the class and feature values for each training example and outputs a classification model, expressed as an ordered set of if-then rules which can be used to classify unseen data. RIPPER also provides methods for obtaining cross-validated estimates of the ruleset's likely performance on unseen data.

In this paper we discuss the predictability of three classes: CA-based and WA-based misrecognition, as well as corrections. As in our earlier studies [5, 3], our feature set includes prosodic features, features based on the ASR inputs and outputs, and features representing the experimental conditions; these features represent properties of the current utterance, the two previous utterances, and the global dialogue history. A full discussion of these features is provided in [5].

In this paper, we evaluate a new set of features based upon some of the turn categories that we have previously predicted — i.e., upon some of our previous dependent variables. We wanted to discover whether classification of a subsequent turn as a correction or an 'aware' could improve our ability to predict whether the current turn had been misrecognized, or whether prediction of prior turns as misrecognized or as 'aware' turns could improve our classification of the current turn as a correction or not.

To provide an upper bound for performance improvements using this approach, we first evaluated these features assuming perfect prediction, i.e., from actual observations. In particular, to predict both CA-based and WA-based prediction, we included the following new features: is the next turn an 'aware'; is the next turn a correction; is the turn following the next turn an 'aware'; is the turn following the next a correction; and is the current turn itself an 'aware'. To predict corrections, we added the following new features to our correction prediction experiments: is the previous turn an 'aware'; is the previous turn misrecognized (in terms of CA or WA); is the turn before the previous turn an 'aware'; is the turn before the previous turn misrecognized (again, in terms of CA or WA); and is the current turn an 'aware'. We considered only a window of two turns after misrecognized and two turns before corrections since, from our data, this window would capture the vast majority of turns in the categories of interest. Two thirds (67%) of corrections occur within two turns of the misrecognized (in terms of CA) turn they correct and 84% of 'awares'.

if (nextaware = T) $\wedge$ (timedur $\geq$ 1.31) **then** *F* (406/14)
if (nextaware = T) $\wedge$ (asrconf $\leq$ -4.34) **then** *F* (112/16)
if (asrconf $\leq$ -2.71) $\wedge$ (nextaware = T) **then** *F* (67/27)
if (asrconf leq -3.73) $\wedge$ (pmnwordsstr $\geq$ 2.43) $\wedge$ (pmnsyls $\leq$ 5.33) $\wedge$ (ppreppau $\leq$ 1.49) **then** *F* (27/5) (tempo $\leq$ 0.88) $\wedge$ (temponorm1 $\leq$ 0.05) **then** *F* (18/1) (asrconf $\leq$ -2.71) $\wedge$ (rejbool=T) **then** *F* (8/1) (asrconf $\leq$ -2.25) $\wedge$ (tempo $\leq$ 1.39) $\wedge$ (syls $\geq$ 3) **then** *F* (12/5) (nextaware=T) $\wedge$ (predur $\leq$ 0.70) **then** *F* (8/1) **else** *T* (1548/52)

Figure 2: *Learned rules for predicting CA-based correct recognition.*

Table 1 presents the best result for each of the three classes that we are are predicting, comparing results from our earlier experiments with our new experiments including prior and subsequent turn classes as features. Adding aware and correction labels appears to improve our ability to identify CA-defined errors. The new predicted error of 7.22%, when we include information about whether the next turn is an 'aware' or not, is significantly better than our best earlier score of 10.48 [5]. However, adding information about whether the following turn is an 'aware' or whether the current turn is an 'aware', or whether either of the next two turns is a correction or an 'aware', does not significantly improve our ability to predict WA-defined misrecognition. However, the actual estimated error (7.60%) when we add all these features is still lower than our previous best result. Our best predicting turn class features do not, however, significantly improved our ability to predict WA-based misrecognitions; although the estimated error (7.60%) is slighted lower than our previous best results (8.64%). This is not surprising, since only CA-based errors even come to the attention of users. For correction prediction, adding information about whether the two preceding turns represent 'awares' or WA-based misrecognitions, or whether the current turn is itself an 'aware', again yields a lower estimated error than any of our previous attempts to predict corrections (13.78%). However, this estimate is not significantly different from our previous best result (15.72%).

Fig 2 presents the rules learned for predicting CA-based misrecognition using our original features plus just the additional information as to whether the next turn is an 'aware' turn — our best performing feature set to date for predicting CA-based misrecognition. Note that the top three best rules do indeed make use of the 'nextaware' feature. Other useful features in the rules are similar to those we have found important for pre-

Turn		Turn ID	Aware of	Corr of	Misrecognition Response Labels
S:	*How may I help you?*				
U:	Get me the train to New York City from Baltimore on Sunday at eight thirty pm	1158	NA	NA	
S:	*Which city do you want to go to?*				
U:	New York City	1159	1158	1158	aware, correction
S:	*Which city do you want to leave from?*				
U:	Baltimore	1160	1158	1158	aware, correction
S:	*Do you want me to find the trains from Baltimore to New York City today at anytime now?*				
U:	No	1161	1158	NA	aware
S:	*How may I help you?*				
U:	Get me the train to New York City from Baltimore on Sunday at eight thirty pm	1162	NA	1158	correction
S:	*Do you want me to find the trains from Baltimore to New York City today at anytime now?*				
U:	No	1163	1162	NA	aware
S:	*How may I help you?*				
U:	Get me the train on Sunday at eight thirty pm	1164	NA	1162	correction

Figure 1: *Dialogue fragment with correction and aware labels.*

Table 1: *Predicting Misrecognitions and Corrections: best-performing feature-sets with and without features derived from other turn types.*

Predicted Class	Error w/o Turn Features	Error w/ Turn Features	Turn Features Used
CA-based misrecognition	10.48 +/- 0.57	7.22% +/- 0.37	nextaware
WA-based misrecognition	8.64 +/- .53	7.60% +/- 0.47	nextaware,nextcorr
Corrections	15.72 +/1 0.80	13.78% +/- 0.82	prevaware,2prevaware,prevwa,2prevwa,curaware

diction in the past, including the length in seconds (timedur) and syllables (syls), speaking rate (tempo) and normalized speaking rate (temponorm1) of the current turn; whether that turn was rejected or not (rejbool) and its ASR acoustic confidence score; the length in seconds of the preceding turn (predur); the length of pause preceding the turn two turns before the current turn (ppreppau); and the mean number of words (pmnwordsstr) and syllables (pmnsyls) per turn calculated over all prior turns. Interestingly, the top four rules, and five of the rules in this ruleset, include features of the current turn's context.

Now we examine how the different categories of new information about the status of prior and subsequent turns as 'awares', corrections, or CA-based or WA-based misrecognitions affect our ability to predict prior misrecognitions or following corrections. We see from Table 2 that, for predicting misrecognitions in general, information about whether the next turn is an 'aware' or not is the single best predictor of prior error.

When such information is absent (e.g. for the feature sets that include only information about whether the next turns are corrections), prediction is significantly poorer. Similarly, information about whether a prior turn is misrecognized appears to be a poor predictor for whether or not the current turn is a correction. Feature sets containing only such information, whether based upon concept or transcription accuracy (prevca, 2 prevca, prevwa, 2prevwa) rank at the bottom in performance of feature sets predicting correction. From our prior experiments [3] we have found that 'awares' can be predicted with some accuracy (12.20% +/- 0.61% error over a baseline of 30% error) from prosodic and other features. So our next step is to see whether this potentially useful information for identifying misrecognitions and corrections can itself be predicted accurately enough to improve prediction when hand-annotated information is absent.

We next compare the performance of rules trained on a training corpus when tested on a test set containing real observations with performance of these rules when tested on a test set containing predicted values for the same turn-based features. For this purpose we divided our 2328 turns into training and test sets, randomly selecting one of the four tasks performed by each subject for testing. This division produced an 1874 turn training set and a 454 turn test set.

Preliminary investigations of the use of predicted turn class features for the prediction of CA-based misrecognitions suggest, however, that we are not yet able to predict the classes from which the features are derived with sufficient accuracy to improve misrecognition prediction, as the observed values do. Table 3 compares results of predicting CA-based misrecognitions from observation-derived turn class features vs. predicted features; note that results for the observed feature prediction are somewhat higher than those presented for the feature sets in Table 2 since the rules are those for the best performing ruleset, i.e., not cross-validated. Neither the best-performing feature set from Table 2, using only information about whether the next turn is an 'aware' or not, nor the entire set of turn class based features, improves our prediction accuracy when these turn classes are predicted, rather than observed. Nor do they approach the cross-validated estimates of Table 2 for CA-based misrecognition prediction. If we examine the noise in the predicted class-based features, we find a likely explanation. The error on the test set for the prediction of the 'aware' class, which was used to generate the features 'nextaware' and '2nextaware'

Feature-Set	Estimated Error
CA-Based Misrecognition	
nextaware	7.22% (+- 0.40%)
nextaware, nextcorr	7.30% (+- 0.37%)
nextaware, 2nextaware	7.30% (+- 0.54%)
nextaware, 2nextaware, nextcorr, 2nextcorr, curaware	7.39% (+- 0.51%)
nextcorr, 2nextcorr	10.52% (+- 0.51%)
nextcorr	10.99% (+- 0.57%)
WA-Based Misrecognition	
nextaware, nextcorr	7.60% (+/- 0.47%)
nextaware, 2nextaware, nextcorr, 2nextcorr, curaware	7.73% (+/- 0.51%)
nextaware	7.98% (+/- 0.51%)
nextaware, 2nextaware	7.98% (+/- 0.51%)
nextcorr	10.48% (+/- 0.52%)
nextcorr, 2nextcorr	11.13% (+/- 0.69%)
Corrections	
prevaware, 2prevaware, prevwa, 2prevwa, curaware	13.78% (+/- 0.82%)
prevaware, 2prevaware, prevca, 2prevca, curaware	14.08% (+/- 0.72%)
prevaware, prevca	14.95% (+/- 0.66%)
prevaware, 2prevaware	14.99% (+/- 0.57%)
prevaware, k25	14.99% (+/- 0.57%)
prevca, 2prevca	15.51% (+/- 0.64%)
prevaware, prevwa	15.85% (+/- 0.79%)
prevca	16.53% (+/- 0.80%)
prevwa	16.70% (+/- 1.04%)
prevwa, 2prevwa	16.87% (+/- 0.70%)

Table 2: *Performance of Turn-Based Feature Sets on Class Predictions*

Feature-Set	Estimated Error
Observed Feature Values	
nextaware	6.61% +/- 1.17%
nextaware, 2nextaware, nextcorr 2nextcorr, curaware	6.61% +/- 1.17%
Predicted Values	
nextaware	14.54% +/- 1.66%
nextaware, 2nextaware, nextcorr 2nextcorr, curaware	14.54% +/- 1.66%

Table 3: *Performance of Predicted vs. Observed Turn-Based Feature Sets on CA-based Misrecognition Prediction*

was 14%, and that for the prediction of corrections was 17%. Both of these are somewhat larger than our cross-validated error estimates of earlier experiments, of 12% and 16%, but even these error rates would probably hurt performance.

Using machine learning techniques more robust to noise, or training prediction rules on predicted values rather than observations for the class-based features (using a separate training set to predict the class features), or finding new features which improve our prediction in particular of 'aware' turns are all measures which might improve our ability to use class-based features to improve the prediction of other classes.

4. Conclusions

In this paper we present results of experiments which explore the utility of class-based turn features — whether subsequent turns are 'aware' sites or corrections and whether previous turns are 'aware' sites or misrecognitions — to predict misrecognitions and corrections in spoken dialogue system. Our experiments have demonstrated that performance can be improved at least in the prediction of concept-based (CA) recognition accuracy by including such information, if the information is accurate. They have also shown that 'aware' site information is of more value in predicting both misrecognitions and corrections than are other class-based turn features. However, further exploration of the use of predicted class values in place of observed values indicates that errors in that prediction seriously degrade performance. In future we will continue these experiments to discover how such errors themselves and how their effect on prediction can be minimized.

5. References

[1] Diane J. Litman, Julia B. Hirschberg, and Marc Swerts, "Predicting automatic speech recognition performance using prosodic cues," in *Proceedings of NAACL-00*, Seattle, May 2000.

[2] Julia B. Hirschberg, Diane J. Litman, and Marc Swerts, "Generalizing prosodic prediction of speech recognition errors," in *Proceedings of ICSLP-00*, Beijing, 2000.

[3] D. Litman, J. Hirschberg, and M. Swerts, "Predicting user reactions to system error," in *Proceedings of ACL-2001*, Toulouse, 2001.

[4] M. Swerts, D. Litman, and J. Hirschberg, "Corrections in spoken dialogue systems," in *Proceedings of ICSLP-00*, Beijing, 2000.

[5] J. Hirschberg, D. Litman, and M. Swerts, "Identifying user corrections automatically in spoken dialogue systems," in *Proceedings of NAACL-01*, Pittsburgh, 2001.

[6] E. Krahmer, M. Swerts, M. Theune, and M. Weegels, "Error spotting in human-machine interactions," in *Proceedings of EUROSPEECH-99*, 1999.

[7] Diane J. Litman and Shimei Pan, "Empirically evaluating an adaptable spoken dialogue system," in *Proceedings of the 7th International Conference on User Modeling (UM)*, 1999.

[8] William Cohen, "Learning trees and rules with set-valued features," in *14th Conference of the American Association of Artificial Intelligence, AAAI*, 1996.