Predicting Barge-in Utterance Errors by using Implicitly Supervised ASR Accuracy and Barge-in Rate per User

Kazunori Komatani
Graduate School of Informatics
Kyoto University
Yoshida, Sakyo, Kyoto 606-8501, Japan
komatani@i.kyoto-u.ac.jp

Alexander I. Rudnicky
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213, U.S.A.
air@cs.cmu.edu

Abstract

Modeling of individual users is a promising way of improving the performance of spoken dialogue systems deployed for the general public and utilized repeatedly. We define "implicitly-supervised" ASR accuracy per user on the basis of responses following the system's explicit confirmations. We combine the estimated ASR accuracy with the user's barge-in rate, which represents how well the user is accustomed to using the system, to predict interpretation errors in barge-in utterances. Experimental results showed that the estimated ASR accuracy improved prediction performance. Since this ASR accuracy and the barge-in rate are obtainable at runtime, they improve prediction performance without the need for manual labeling.

1 Introduction

The automatic speech recognition (ASR) result is the most important input information for spoken dialogue systems, and therefore, its errors are critical problems. Many researchers have tackled this problem by developing ASR confidence measures based on utterance-level information and dialogue-level information (Litman et al., 1999; Walker et al., 2000). Especially in systems deployed for the general public such as those of (Komatani et al., 2005) and (Raux et al., 2006), the systems need to correctly detect interpretation errors caused by various utterances made by various kinds of users including novices. Furthermore, since some users access such systems repeatedly (Komatani et al., 2007), error detection by using individual user models would be a promising way of improving performance.

In another aspect in dialogue systems, certain dialogue patterns indicate that ASR results in certain positions are reliable. For example, Sudoh and Nakano (2005) proposed "post-dialogue confidence scoring" in which ASR results corresponding to the user's intention upon dialogue completion are assumed to be correct and are used for confidence scoring. Bohus and Rudnicky (2007) proposed "implicitly-supervised learning" in which users' responses following the system's explicit confirmations are used for confidence scoring. If ASR results can be regarded as reliable after the dialogue, machine learning algorithms can use such ASR results as teacher signals. This approach enables the system to improve its performance without any manual labeling or transcription, a task which requires much time and labor when spoken dialogue systems are developed.

We focus on users' affirmative and negative responses to the system's explicit confirmations as in (Bohus and Rudnicky, 2007) and estimate the user's ASR accuracy on the basis of his or her history of responses. The estimated ASR accuracy is combined with the user's barge-in rate to predict the interpretation error in the current barge-in utterance. Because the estimated ASR accuracy and the barge-in rate per user are obtainable at runtime, it is possible to improve prediction performance without any manual transcription or labeling.

2 Implicitly Supervised Estimation of ASR Accuracy

2.1 Predicting Errors in Barge-in Utterance

We aim to predict interpretation errors in barge-in utterances at runtime. These errors are caused by ASR errors, and barge-in utterances are more prone to be misrecognized. A user study conducted by Rose and Kim (2003) revealed that there are many more disfluencies when users barge-in compared with when users wait until the system prompt ends. It is difficult to select the erroneous utterances to be rejected by using a classifier that

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 89–92,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

distinguishes speech from noise on the basis of the Gaussian Mixture Model (Lee et al., 2004); such disfluencies and resulting utterance fragments are parts of human speech.

Barge-in utterances are, therefore, more difficult to recognize correctly, especially when novice users barge-in. To detect their interpretation errors, other features should be incorporated instead of speech signals or ASR results. We predicted the interpretation errors in barge-in utterances on the basis of each user's barge-in rate (Komatani et al., 2008). This rate intuitively corresponds to how well users are accustomed to using the system, especially to its barge-in function.

Furthermore, we utilize a user's ASR accuracy in his or her history of all utterances including barge-ins. The ASR accuracy also indicates the user's habituation. However, it has been shown that the user's ASR accuracy and barge-in rate do not improve simultaneously (Komatani et al., 2007). In fact, some expert users have low barge-in rates. We thus can predict whether a barge-in utterance will be correctly interpreted or not by integrating the user's current ASR accuracy and barge-in rate.

2.2 Estimating ASR Accuracy by using Implicitly Supervised Labels

To perform runtime prediction, we use information derived from the dialogue patterns to estimate the user's ASR accuracy. We estimate the accuracy on the basis of the user's history of responses following the system's explicit confirmations such as "Leaving from Kyoto Station. Is that correct?"

Specifically, we assume that the ASR results of affirmative or negative responses following explicit confirmations are correct and that the user utterances corresponding to the content of the affirmative responses are also correct. We further assume that the remaining utterances are incorrect because users do not often respond with "no" for explicit confirmations containing incorrect content and instead repeat their original utterances. Consequently, we regard that the ASR results of the following utterances are correct: (1) affirmative responses and their immediately preceding utterances and (2) negative responses. Accordingly, all other utterances are incorrect. We thus calculate the user's estimated ASR accuracy by using the user's utterance history, as follows:

$$\begin{aligned}&(\text{Estimated ASR accuracy})\\&\quad= \frac{2 \times (\#\text{affirmatives}) + (\#\text{negatives})}{(\#\text{all utterances})} \qquad (1)\end{aligned}$$

2.3 Predicting Errors by Using Barge-in Rate and ASR Accuracy

We predict the errors in barge-in utterances by using a logistic regression function:

$$P = \frac{1}{1 + \exp(-(a_1 x_1 + a_2 x_2 + b))}.$$

Its inputs x_1 and x_2 are the barge-in rate until the current utterance and ASR accuracy until the previous utterance. To account for temporal changes in barge-in rates, we set a window when calculating them (Komatani et al., 2008). That is, when the window width is N, the rates are calculated by using only the last N utterances, and the previous utterances are discarded. When the window width exceeds the total number of utterances by the user, the barge-in rates are calculated by using all the user's utterances. Thus, when the width exceeds 2,838, the maximum number of utterances made by one user in our data, the barge-in rates equal the average rates of all previous utterances by the user.

We calculate the estimated ASR accuracy every time a user makes an affirmative or negative response. When the user makes other utterances, we take the estimated accuracy when the *last* affirmative/negative response is made to be the accuracy of those utterances.

3 Experimental Evaluation

3.1 Target Data

We used data collected by the Kyoto City Bus Information System (Komatani et al., 2005). This system locates a bus that a user wants to ride and tells the user how long it will be before the bus arrives. The system was accessible to the public by telephone. It used the safest strategy to prevent erroneous responses, that is, to make explicit confirmations for all ASR results.

We used 27,519 utterances after removing calls whose phone numbers were not recorded and those the system developer called for debugging. From that number, there were 7,193 barge-in utterances, i.e., utterances that a user starts speaking during a system prompt. The phone numbers of the calls were recorded, and we assumed that each

Table 1: ASR accuracy by response type

	Correct	Incorrect	Total	(Acc.)
Affirmative	9,055	246	9,301	(97.4%)
Negative	2,006	289	2,295	(87.4%)
Other	8,914	7,009	15,923	(57.9%)
Total	19,975	7,544	27,519	(72.6%)

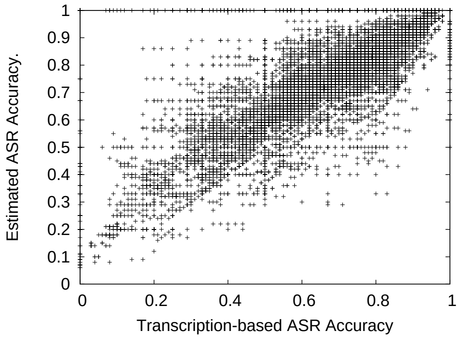

Figure 1: Correlation between transcription-based and estimated ASR accuracy

number corresponded to one individual. Most of the numbers were those of mobile phones, which are usually not shared, so the assumption seems reasonable.

Each utterance was transcribed and its interpretation result, correct or not, was given manually. We assumed that an interpretation result for an utterance was correct if all content words in its transcription were correctly included in the result. The result was regarded as an error if any content words were missed or misrecognized.

3.2 Verifying Implicitly Supervised Labels

We confirmed our assumption that the ASR results of affirmative or negative responses following explicit confirmations are correct. We classified the user utterances into affirmatives, negatives, and other, and calculated the ASR accuracies (precision rates) as shown in Table 1. Affirmatives include *hai* ('yes'), *soudesu* ('that's right'), OK, etc; and negatives include *iie* ('no'), *chigaimasu* ('I don't agree'), *dame* ('No good'), etc. The table indicates that the ASR accuracies of affirmatives and negatives were high. One of the reasons for the high accuracy was that these utterances are much shorter than other content words, so they were not confused with other content words. Another reason was that the system often gave help messages such as "Please answer *yes* or *no*."

We then analyzed the correlation between the transcription-based ASR accuracy and the estimated ASR accuracy based on Equation 1. We plotted the two ASR accuracies in Figure 1 for 26,231 utterances made after at least one affirmative/negative response by the user. The correlation coefficient between them was 0.806. Although the assumption that all ASR results of affirmative/negative responses are correct might be strong, the estimated ASR accuracy had a high correlation with the transcription-based ASR accuracy.

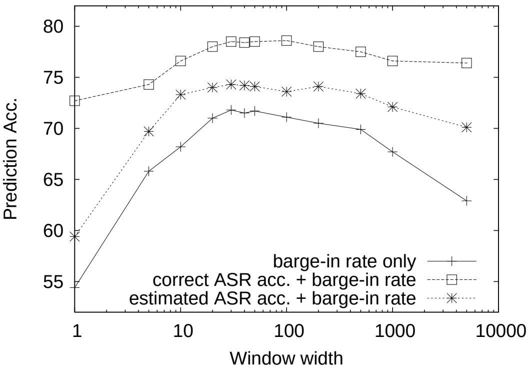

Figure 2: Prediction accuracy with various window widths

3.3 Prediction using Implicitly Supervised Labels

We measured the prediction accuracy for 7,193 barge-in utterances under several conditions. We did not set windows when calculating the ASR accuracies and thus used all previous utterances of the user, because the windows did not improve prediction accuracy. One of the reasons for this lack of improvement is that the ASR accuracies did not change as significantly as the barge-in rates because the accuracies of frequent users converged earlier (Komatani et al., 2007).

We first confirmed the effect of the transcription-based ("correct", hereafter) ASR accuracy. As shown in Figure 2 and Table 2, the prediction accuracy improved by using the ASR accuracy in addition to the barge-in rate. The best prediction accuracy (78.6%) was when the window width of the barge-in rate was 100, and the accuracy converged when the width was 30. The prediction accuracy was 72.7% when only the "correct" ASR accuracy was used, and the prediction accuracy was 71.8% when only the barge-in rate was used. Thus, the prediction accuracy was better when both inputs were used rather than when either input was used. This

Table 2: Best prediction accuracies for each condition and window width w

Conditions (Used inputs)	Prediction acc. (%)
barge-in rate	71.8 (w=30)
correct ASR acc.	72.7
+ barge-in rate	78.6 (w=100)
estimated ASR acc.	59.4
+ barge-in rate	74.3 (w=30)

fact indicates that both the barge-in rate and ASR accuracy have different information and contribute to the prediction accuracy.

Next, we analyzed the prediction accuracy after replacing the correct ASR accuracy with the estimated one described in Section 2.2. The best accuracy (74.3%) was when the window width was 30. This accuracy was higher than that of using only barge-in rates. Hence, the estimated ASR accuracy without manual labeling is effective in predicting the errors in barge-in utterances at runtime.

4 Conclusion

We proposed a method to estimate the errors in barge-in utterances by using a novel dialogue-level feature obtainable at runtime. This method does not require supervised manual labeling. The estimated ASR accuracy based on the user's utterance history was dependable in predicting the errors in the current utterance. We thus showed that ASR accuracy can be estimated in an implicitly supervised manner.

The information obtained by our method can be used for confidence scoring. Thus, our future work will include integrating the proposed features with bottom-up information such as acoustic-score-based confidence measures. Additionally, we simply assumed in this study that all affirmative and negative responses following the explicit confirmation are correct. By modeling this assumption more precisely, prediction accuracy will improve. Finally, we identified individuals on the basis of their telephone numbers. If we utilize user identification techniques to account for situations when no speaker information is available beforehand, this method can be applied to systems other than telephone-based ones, e.g., to human-robot interaction.

Acknowledgments

We are grateful to Prof. Tatsuya Kawahara of Kyoto University who led the project of the Kyoto City Bus Information System.

References

Dan Bohus and Alexander Rudnicky. 2007. Implicitly-supervised learning in spoken language interfaces: an application to the confidence annotation problem. In *Proc. SIGdial Workshop on Discourse and Dialogue*, pages 256–264.

Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. 2005. User modeling in spoken dialogue systems to generate flexible guidance. *User Modeling and User-Adapted Interaction*, 15(1):169–183.

Kazunori Komatani, Tatsuya Kawahara, and Hiroshi G. Okuno. 2007. Analyzing temporal transition of real user's behaviors in a spoken dialogue system. In *Proc. INTERSPEECH*, pages 142–145.

Kazunori Komatani, Tatsuya Kawahara, and Hiroshi G. Okuno. 2008. Predicting ASR errors by exploiting barge-in rate of individual users for spoken dialogue systems. In *Proc. INTERSPEECH*, pages 183–186.

Akinobu Lee, Keisuke Nakamura, Ryuichi Nisimura, Hiroshi Saruwatari, and Kiyohiro Shikano. 2004. Noice robust real world spoken dialogue system using GMM based rejection of unintended inputs. In *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, pages 173–176.

Diane J. Litman, Marilyn A. Walker, and Michael S. Kearns. 1999. Automatic detection of poor speech recognition at the dialogue level. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 309–316.

Antoine Raux, Dan Bohus, Brian Langner, Alan W. Black, and Maxine Eskenazi. 2006. Doing research on a deployed spoken dialogue system: One year of Let's Go! experience. In *Proc. INTERSPEECH*.

Richard C. Rose and Hong Kook Kim. 2003. A hybrid barge-in procedure for more reliable turn-taking in human-machine dialog systems. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 198–203.

Katsuhito Sudoh and Mikio Nanano. 2005. Post-dialogue confidence scoring for unsupervised statistical language model training. *Speech Communication*, 45:387–400.

Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman. 2000. Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You? In *Proc. North American Chapter of Association for Computational Linguistics (NAACL)*, pages 210–217.