LEARNING DECISION TREES TO DETERMINE TURN-TAKING BY SPOKEN DIALOGUE SYSTEMS

Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, Kiyoaki Aikawa

NTT Communication Science Laboratories
NTT Corporation
3-1 Morinosato-Wakamiya
Atsugi, Kanagawa 243-0198, Japan
{sato, rh, nakano}@atom.brl.ntt.co.jp, {tamoto, aik}@idea.brl.ntt.co.jp

ABSTRACT

This paper presents a method for deciding the timing of turn-taking in spoken dialogue systems. This method uses a decision tree learned from the corpus of dialogues between human users and systems in which desirable turn-taking behaviors are annotated by hand. It utilizes a variety of attributes, such as recognition and understanding results and prosodic information. Unlike most of the existing systems it enables spoken dialogue systems to decide the timing of turn-taking based on not only pauses but also other features, so that users can speak to the system even if they put pauses in the middle of their utterances. The result of a preliminary experiment shows that the learned decision tree outperforms the baseline strategy, which takes turn at every user pauses.

1. INTRODUCTION

Recent advances in speech and language technologies have made it possible to build computer systems that can communicate with humans using spoken dialogue. Most of these systems take a dialogue turn when they detect a pause longer than a certain length and release the turn when they finish speaking. This prevents the user from speaking naturally to them; users must be careful not to pause before they finish telling the system everything they want to. The system sometimes interrupts during a pause in the middle of user utterance. Lengthening the minimum allowable pause might solve the problem. In that case, however, the system cannot respond immediately when the user finish speaking. It is therefore necessary to find a way to decide the timing of the turn-taking based on not only pauses but also other features.

The objective of this research is to determine when the system should take a turn. The human-human dialogue corpora have been studied to investigate the timing of turn-taking and backchannels. Note that, in this paper, we consider making a backchannel does not take a turn. However, little work has been done on turn-taking of spoken dialogue systems. We have been studying this problem and have developed a system that takes turns based on understanding results as well as pauses [1]. The turn-taking strategy of this system, however, was devised based on the developer's intuition, not on a corpus-based study.

This paper proposes a method for learning an algorithm for deciding turn-taking from corpora of dialogues between human users and spoken dialogue systems. User pauses are considered appropriate places for the system to take a turn, and the learned algorithm distinguishes turn-taking pauses from others. We use decision tree learning in our approach because many kinds of features, such as recognition results, understanding results, and prosodic features, are considered to play roles in turn-taking; decision tree learning is a suitable way to combine them to form a classifier. Results of preliminary experiments using a corpus of dialogues between human naive users and a meeting room reservation system suggests the usefulness of our approach.

2. RELATED WORK

There have been a couple of works on turn-taking in human-human dialogues. Koiso *et al.* analyzed human-human dialogues and suggested that humans use syntactic and prosodic information as turn-taking cues [2]. Tamoto and Kawabata examined turn-taking cues by analyzing human-human dialogues [3]. However, human-computer dialogues differ from human-human dialogues in many respects [4]. Moreover, in the case of spoken dialogue systems, speech recognition and language understanding errors are inevitable. The results of human-human dialogue analyses are not always applicable to human-computer dialogues in the same way, though we can make reference to those results.

As for spoken dialogue system research, Bell *et al.* built a spoken dialogue system that decides the timing of turn-taking based on syntactic information cues [5]. They suggest that expressions at the end of sentences are useful. However, they do not consider recognition error and did not evaluate the system.

3. APPROACH

3.1. Overview

We consider user pauses, or the end of user utterances, as being an appropriate time to take turn. This is because an experiment suggested that some users felt unpleasant when the system barged in during their utterances [6]. This is not to deny the possibility that, in some cases, the system's barge-in utterances are appropriate. The challenge is, therefore, to develop an algorithm that lets the system determine whether it should take a turn when the user pauses. In their human-human dialogue study, Koiso et al. [2] report that prosodic features, contexts, syntactic information,

Response of two labelers	The number of utterances
Turn-taking, both labelers	2,503
Different, each other	307
Backchannel, both labelers	1,435

Table 1. Agreement between two labelers.

Agreement between the recording system system and the labelers	The number of utterances
Agreement	3,001
Disagreement	937

Table 2. Agreement between correct labels and the system response.

and dialogue history are related to the timing of turn-taking. Although not all of these features are available in spoken dialogue systems and some other features may be effective, it is reasonable to consider that many kinds of features also need to be utilized for determining the timing of turn-taking in spoken dialogue systems. These include both symbolic and numeric features. We therefore use a decision tree learning method [7], so that both symbolic and numeric features can be dealt with. Decision trees are learned from the data of dialogues between human users and the system.

3.2. Classes

The classes to be output by the learned decision tree are one of two behaviors: *taking a turn* and *not taking a turn*. The correct classes, which we call reference behaviors, must be labeled to the system logs of human-computer dialogues for the decision tree to be learned. This is done by human subjects, who have not engaged in the dialogues.

3.3. Features

We decided that several kinds of features should be used in the decision tree learning. Needless to say, the user speech interval just before the pause at which the system decides turn-taking is influential on the system's appropriate turn-taking behavior. We refer the interval simply as *user utterance* when it is clear from the context.

An analysis of human-human dialogues has shown that linguistic information, especially sentence-final expressions, plays an important role in turn-taking, even if there are no user pauses [2]. This suggests that speech recognition results on user utterances can be used for deciding turn-taking in spoken dialogue systems. It is also suggested that the prosodic features of the utterances are related to the timing of turn-taking. In addition, it is reasonable to consider the duration of the utterances and the number of words in its recognition result because the timing of turn-taking could be related to the amount and content of the information conveyed to the system. We also think that the partial results of the user utterance understanding up to that point of time [1] is effective in turn-taking.

	Accuracy(%)
Proposed method (evaluated on test sets)	83.9
Proposed method (predicted)	86.6
Recording System	76.2
Baseline	63.7

Table 3. Accuracy of the learned decision tree obtained by cross-validation.

4. EXPERIMENT

This section reports the details and results of a preliminary experiment conducted to examine the effectiveness of our approach.

4.1. Dialogue Data Collection

We recorded dialogues between humans and a Japanese spoken dialogue system to collect training and test data. The system was built with the WIT spoken dialogue system toolkit [8] and employers Julius [9] as the speech recognizer and NTT Cyber Space Lab's Final Fluet [10] as the speech synthesizer. This system regards silence longer than 0.75 seconds as a user pause. The dialogue domain is meeting room reservation. Vocabulary size of the speech recognizer was 161. The tasks were to reserve two different rooms on the same date, and reserving a room on two different dates. Subjects were given either task and shown the date, start time, end time, and name of the conference room. Although these tasks seem simple, we chose them for the preliminary experiment because we suspected that, if the tasks were too complicated, the subjects might only make short utterances because of the poor speech recognition accuracy due to the limited amount of the language model training data.

An understanding state of the spoken dialogue system is represented by a frame. The system is able to understand user utterances incrementally so that the understanding result is obtained at each pause. The system decides whether it should take a turn or not by referring to its understanding state. Turn-taking by the system is decided according to the following heuristics.

> If a user says a phrase that means "I'd like to reserve", the system takes a turn and asks for the values of all slots that are empty. If all slots are filled, the system takes a turn for confirmation. In other situations, the system utters only backchannels.

We call the system a recording system in the following discussion.

We collected 210 dialogues and totally 4,768 user utterances. Twenty-four subjects, ranging from 19 to 35 years old, took part in the recording. The gender ratio was 1:1. We recorded about ten dialogues per subject. In 69% of the dialogues, the task was achieved within five minutes.

4.2. Hand Labeling of Reference Behaviors

Next, labeling of correct classifications was done. Non-experts, whom we call labelers, listened to the recorded dialogues and put turn-taking labels on pauses where they thought the system should take a turn. They put backchannel labels on pauses where they thought the system should utter a backchannel or do nothing. The

sound stopped when system detected the end of user utterances, which were almost pauses. When the sound stopped, labelers placed labels. The result of these selections were used for machine learning as correct classes. Each dialogue was labeled by two labelers. The distribution of labels placed by that two labelers, is shown in Table 1. The agreement was 92.8%. This indicates that in most cases human labelers agree on preferable turn-taking behaviors, and their labels can be used as references.

Only when two labels agree, were they used as correct classes. We did not use other data for both training and testing. The numbers of agreement labels are shown in Table 2. The turn-taking accuracy of the recording system was 76.2%.

4.3. Features

In this experiment, we used 114 features in the following categories.

Categories: The syntactic and semantic categories. If at least one word in the recognition result for user utterances falls in a category, say c, the value of the feature c is "yes", and otherwise "no". Function words are classified into small categories. We use 30 category features.

Final word category: The syntactic and semantic categories of the final word of the recognition result of the utterance. This is represented by 30 features, each of which corresponds to a syntactic and semantic category above.

Number of words: The number of words in the recognition result of the user utterance.

Understanding state: The content of system's understanding state (or understanding result) after understanding the user utterance. Each feature in this category corresponds to each slot in the frames representing the understanding state. We use eight slots. The value of each feature is either "empty" or "non-empty", depending on whether the corresponding slot value is empty or not.

Understanding state change: Change in the value of each slot in the understanding state after the user utterance.

Duration: Duration of user utterances.

Prosodic features: Pitch and power parameters used in Noguchi and Den's study on finding the context of backchannels [11]. They used 17 pitch parameters. We use the same 17 and one confidence value. The confidence value depends on the length of user utterance, because it is difficult to detect pitch in the case of a short utterance. The same parameters are useful with respect to power; thus there are 18 power parameters.

4.4. Evaluation of Machine Learning

We made a decision tree based on the previous features obtained by machine learning, and compared our method with other heuristics to evaluate it. We used the C4.5 algorithm to make the decision tree [7]. We used it with the default values for all options. To evaluate the accuracy of the learned decision tree, we performed 10-fold cross validation. The result is shown in Table 3. The accuracy was computed with a pruned tree.

Feature set	Actual accuracy on test sets (%)	Predicted accuracy (%)
Full set	83.9	86.6
without *categories*	81.8	85.2
without *final word category*	84.2	86.8
without *number of words*	84.9	86.8
without *understanding state*	78.7	82.2
without *understanding state change*	84.6	86.8
without *duration*	84.4	86.9
without *prosody*	85.5	86.7
categories and *understanding state* only	85.2	85.6
understanding state only	76.4	75.1

Table 4. Change in performances of learned decision trees when some features were not used.

If the system always took a turn at the end of a user utterance, we call it a baseline algorithm. The turn-taking accuracy rate was 63.7%. The turn-taking accuracy for the recording system was 76.2%. On the other hand, that of our method was 83.9%. To estimate the effectiveness of the features in each category, we investigated the change in the accuracy when each category of features was not used. The results are shown in Table 4. Since they show that the *understanding state* features as well as the *categories* features are crucial for the system to decide turn-taking, we also did learning experiments only with these features, the results of which are also shown in Table 4. These results show that using both the *understanding state* features and the *categories* features does not degrade the accuracy very much, whereas using only the *understanding state* features is insufficient.

Table 5 shows some of the paths from the root to a leaf of the tree built with the *categories* and the *understanding state* features. These paths are the highest coverage ones. In the table, a feature whose name ends with *slot* is an *understanding state* feature and a feature whose name starts with *category* is a *category* feature. For example, the *start_time_slot* feature means whether the value of the slot containing the start time of the requested reservation in the understanding frame is empty or not. The *category_request_aux* feature means whether the recognition result of the user utterance includes the auxiliary word used in verbal phrases expressing requests such as 'shitainodesuga'. Because these features are specific to the task domain, we do not explain them in detail.

5. DISCUSSION

Although there are limitations in this experiment in that the tasks were small and not realistic, these results suggest that this method is useful, because turn-taking accuracy improved compared with the baseline algorithm. They also suggest the possibility that incremental understanding, which outputs the partial result of understanding even before the end of turn is decided, is indispensable.

As Table 5 shows, the obtained tree is highly domain-dependent and difficult to interpret. We suspect that, to obtain simpler and

Path	Behavior	Coverage (%)	Error rate (%)
end_time_slot $\neq$ empty & *discourse_action_slot* $\neq$ empty & *start_time_slot* $\neq$ empty	turn-taking	24.1	7.2
end_time_slot = empty & *category_request_aux* = yes & *category_reserve* = no & *category_particle_made* = no & *start_time_slot* = empty & *category_reservation_possiblity* = no & *room1_slot* = empty	backchannel	14.1	6.6
end_time_slot $\neq$ empty & *discourse_action_slot* = empty & *category_request* = no & *room1_slot* $\neq$ empty & *category_interjection* = no & *category_time2* = empty & *category_reset_exp* = no & *topic_slot* = no & *category_slot* = no & *date1_slot* $\neq$ empty	turn-taking	6.6	20.1

Table 5. Example paths in the decision tree.

more general rules for turn-taking, we need to explore features other than what we have considered. We hope the same experiments in other domains will lead to finding such features.

While the results of this experiment show that the prosodic features are not crucial for this system, we cannot yet conclude that they are not useful in any systems from this preliminary experiment. We need to perform experiments in other task domains in addition to examining the extraction accuracy of prosodic features. However, considering the result of Koiso *et al.*'s study on human-human dialogues, we suspect that only prosodic information is not enough to decide the timing of turn-taking.

6. SUMMARY AND FUTURE WORK

This paper presented a practical method of learning an algorithm for deciding whether a spoken dialogue system should take a turn or not when a user pauses. The result of a preliminary experiment showed our method is effective in the meeting room reservation task we used.

In addition to conducting experiments in other task domains, we plan to incorporate the learned decision trees into the system to evaluate their effectiveness in the dialogues with human naive users by observing the task completion rate and time as well as user satisfaction.

7. ACKNOWLEDGEMENTS

We thank Dr. Hiroshi Murase, the executive manager of the Media Information Laboratory, for his encouragement and comments. We also thank all the members of the Dialogue Understanding Research Group for their help.

8. REFERENCES

[1] M. Nakano, N. Miyazaki, J. Hirasawa, K. Dohsaka, and T. Kawabata, "Understanding unsegmented user utterances in real-time spoken dialogue systems," in *Proc. 37th ACL*, 1999, pp. 200–207.

[2] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs," *Language and Speech*, 41(3–4):295–321, 1998.

[3] M. Tamoto and T. Kawabata, "Analysis of Speaker Transition in Cooperative Task Dialogues," in *IPSJ-SLP,96-SIG-SLP-12-3,Information Processing Society of Japan*, 1996, pp. 13–18, (in Japanese).

[4] C. Doran, J. Aberdeen, L. Damianos, and L. Hirschman, "Comparing several aspects of human-computer and human-human dialogues," in *Proc. Second SIGDial*, 2001, pp. 48–57.

[5] L. Bell, J. Boye, and J. Gustafson, "Real-time handling of fragmented utterances," in *NAACL-2001 Workshop on Adaptation in Dialogue Systems*, 2001.

[6] J. Hirasawa, M. Nakano, T. Kawabata, and K. Aikawa, "Effects of system barge-in responses on user impressions," in *Proc. Eurospeech*, 1999.

[7] J. R. Quinlan, *C4.5: Programs for Machine Learning*, The Morgan Kaufmann, 1992.

[8] M. Nakano, N. Miyazaki, N. Yasuda, A. Sugiyama, J. Hirasawa, K. Dohsaka, and K. Aikawa, "WIT: A toolkit for building robust and real-time spoken dialogue systems," 2000, pp. 150–159.

[9] A. Lee, T. Kawahara, and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine," in *Proc. Eurospeech*, 2001, pp. 1691–1694.

[10] S. Takano, K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima, "A Japanese TTS Syntem Based on Multi-form Units and a Speech Modification Algorithm with Harmonics Reconstruction," *IEEE Transactions on Speech and Processing*, 9(1):3–10, 2001.

[11] H. Noguchi and Y. Den, "Prosody-based detection of the context of backchannel responses," in *Proc. ICSLP*, 1998.