Vol.2019-SLP-128 No.3
2019/7/19

音声対話情報提供システムにおけるシステム応答速度の情報伝達効率への影響の分析

佐伯 真於[1]　松山　洋一[1]　藤江 真也[2,1]　高津 弘明[1]　小林 哲則[1]

概要：In this research, we analyze the effect of system response latency of a spoken dialog system that can efficiently deliver massive amount of information to users. The system assumes that users primarily listen to a certain amount of information, such as news, and sometimes spontaneously ask for clarification and details of the contents. For such an information consumption model, faster response latency should be crucial. We propose a recurrent neural network (RNN) based model that incrementally estimates users' end-of-utterances to achieve faster turn-taking. We conducted a user study comparing a traditional VAD and the proposed RNN model with 67 subjects. The result shows the proposed method did respond faster by an average of 782 ms, and the further analysis shows that system response latency directly affects the efficiency of information transfer (EoIT) for certain types of users.

Analysis of the Effect of Response Latency of Spoken Dialogue System Delivering Massive Amount of Information

Mao Saeki[1]　Yoichi Matsuyama[1]　Fujie Shinya[2,1]　Takatsu Hiroaki[1]　Tetsunori Kobayashi[1]

1. Introduction

In order to determine when the system should begin speech, conventional approach have been to use Voice Activity Detection (VAD) techniques and consider the end-of-utterance as the user's end-of-turn. However this approach has limitations in the speed of system response since it must wait until speech terminates. Analysis of human-human spoken dialog shows that 50% of turn taking are done before the previous speaker terminates speech, and 80%are done under 600ms after the term of speech as we show in this research. In order to achieve this fast response we propose a recurrent neural network (RNN) based model to predict the end-of-turn probability of a user in each frame.

Previously [1] has focused on continuous end-of-utterance prediction for generating back-channels in an attentive listening system, since timing in back-channelling is key to create synchrony in a conversation. However we believe timing to be important for any utterance from the system. [2], [3] has developed similar general end-of-utterance prediction model that do not require lexical information, however its effect on real time dialog system has not been verified.

In this research, we aim to incorporate a general end-of-utterance prediction model in a spoken dialog system to reduce the system's turn-taking speed. We will also like to find the effect of turn-taking speed on the task of large information transfer through a dialog. Therefore we conducted a user experiment to measure the effects.

The rest of the paper is organized as follows. In the next section, we explain the data used for building the end-of-turn prediction model. In section 3, the end-of-

[1] 早稲田大学
Waseda University
[2] 千葉工業大学
Chiba Institute of Technology

turn prediction model is explained. The experiment and its result is explained in section 4, and section 5 gives the conclusion.

2. Data

2.1 Data Outline

The system used in this research is a Japanese spoken dialog system. Therefore for the analysis of response speed and the training of the turn prediction model, we use four Japanese human-human spoken dialog corpus below which has been previously collected, or is openly available.

- Waseda Corpus (W): Collected by the Waseda University and ATR-Trek. Includes a dialog assuming a restaurant search system. 391 dialogues of 16 hours and 30 minutes.
- Japanese Spoken Language Corpus (CSJ): Interview dialog between a lecturer and the interviewer. 58 dialogues of 12 hours and 10 minutes.
- RWCP voice dataset (RWCP): Dialog assumed between a car retailer or a travel agency staff and the customer. 61 dialogues of 8 hours and 29 minutes.
- PASD spoken dialogue dataset (PASD): Dialogue in several tasks, partially including human-machine interaction. 61 dialogues of 5 hours and 8 minutes.

2.2 Analysis of turn-taking

We analyze the timing of turn-taking made in the four corpus aforementioned. To automatically count turn-taking, we use voice activity section. For datasets that include voice activity sections (CSJ, RWCP), those information were used. For the other corpus, openly available tool (https://github.com/ASTL-NICT/VAD) was used to detect the sections. In the case where pause between voice activity sections was under 700ms, the two sections where merged.

Having preprocessed the corpus, we analyzed the time between one speaker ending speech and the other speaker beginning speech. The count of turn-taking for each corpus was 30,872 in W, 8,132 in CSJ 3,493 in RWCP and 4,131 in PASD. Invariant to the difference of tasks, approximately 50% of the turn-taking was done before the former speaker ending speech, 80% was done under 600ms, and 95% was done under 1,000ms. PASD includes human-machine dialogs, and system response speed tends to be slower than humans.

2.3 Labeling of speech-turn retain

In the previous analysis, sections separated by non-voiced sections longer than 700ms was considered to be the speech section, and its interchange was considered a turn-taking. Here, we model the state of retaining speech more precisely by considering the interrelationship of voice active section of two people in a dialog.

First, multiple voice active sections are merged when the non-voiced section between them are shorter than θ_{short_pause}. We call the merged section a **speech section**. Next, utterance shorter than $\theta_{short\_utterance}$ is called a **short utterance**, and consider then to be a speech that do not take a turn.

A speech that is not a short utterance is a speech that takes a turn, and we call this a **turn-taking speech**. A turn-taking speech is merged to ignore pause shorter than θ_{keep_turn}. Here, θ_{keep_turn} is longer than θ_{short_pause}, meaning that once a turn has been taken, it is retained ignoring non-voiced section of some length.

As an exception, if a speech longer than θ_{short_pause} is included in a turn-taking speech (if it begins after and and end before the other speaker's on-going turn-taking speech), we consider this a short utterance. With this rule, any length of speech could be considered a short-utterance, however in a cooperative dialogue that is intended in this research, the speaker who has not taken the turn will not continue to speak. Therefore including such speech in a short pause will have no practical issue.

3. End-of-turn prediction model

3.1 Model outline

The purpose of the model is to predict whether the user has ended his/her turn or not. Of the states we have defined in the previous section, turn-taking speech, short utterance and non-speech are exclusive, and therefore build a model that classify these three states. We also predict whether if it is a voice active section and whether if it is a voice or not at the same time. After training a model to classify all these states, we only use the classification for turn-taking speech to consider whether the user has ended the turn or not.

3.2 Feature Extraction

In order to detect end-of-utterance, prosodoc information such as pitch movement is known to be important[4], and is widely used in many conventional methods. However the precision of pitch extraction can differ depending on the method, and smoothing over several frames is necessary to maintain precision. Therefore pitch extraction can lead to the slowdown of the whole process

which is critical in a dialog system. In the other hand, using MFCC features widely used in speech recognition can eliminate process slowdown, however loses the prosodoc information. In this research, we construct an auto encoder neural network that encodes and decodes a narrow band spectrum, and use the output of the intermediate layer as the acoustic features. Precisely, 10 consequent 256 points of a power spectrum gained every 10ms is used as input, and a 256 dimensional intermediate output is used as the acoustic features.

To predict the state of turn-taking, lexical information is also known to be useful. In this research, we extract lexical features from the partial speech recognition result continuously obtained by the streaming speech recognition of the Google Cloud Speech API (https://cloud.google.com/speech/). Precisely, we use the 512 dimensional intermediate output of the LSTM language model as linguistic features. Since speech recognition result cannot be obtained every frame, a zero vector is used in such frames.

3.3 Speech state prediction model

Using the obtained acoustic and linguistic features as input, a model that continuously predicts the state of turn-taking is constructed using a neural network.

In each frame, the probability of the input being a turn-taking speech, short-utterance or neither (not voiced) is predicted. Also, the probability of the user simply speaking or not, and whether it is a voice or not is predicted at the same time.

4. Experiment

4.1 Turn prediction model training

The auto encoder used for acoustic feature extraction is trained using 10 speech randomly chosen from each of the 1,000 lecture recordings included in the Japanese Spoken Language Corpus. The turn-prediction model is trained using the customer voice of 100 dialogues in the Waseda Corpus. θ_{short_pause} and $\theta_{short\_utterance}$ is set to 1000ms, and θ_{keep_turn} is set to 1500ms.

4.2 Experiment platform

Our proposed model is experimented on a previously developed spoken dialog system that can efficiently deliver massive amount of information to users[5]. The system assumes that users primarily listen to a certain amount of information, and sometimes spontaneously ask for clarification and details of the contents. In this research, we use news as the target of information delivery, hence call the system the News Dialog System (NDS). Users are informed of news articles through an audio headset with no visual display, and may interrupt the system at any time for comment or clarification. The NDS, depending on the user's feedback decides to include, exclude or change the order in which details of the news article is conveyed. The NDS has a end-of-turn detection module which predicts every frame whether the user has finished his/her turn, and our end-of-turn prediction model is incorporated into this module.

Next, we will describe the baseline implementation of this module on which our model can be compared. The baseline implementation uses voice activity detection (VAD) to detect a end of user utterance, and if no other voice activity follow within a threshold time, end-of-turn is predicted. Although this is a simple and faulty approach that leads to a sluggish system response if the threshold is too long, or a frequent interruption of user speech if it is too short, it is a common approach used in many SDS. Along with the VAD, results of the ASR module is used to determine the end-of-turn. The ASR module processes and updates the recognition result incrementally. If the ASR determines that the recognition result cannot be updated anymore, it returns a final signal which the end-of-turn detection module uses to predict end-of-turn. For the ASR module, we used the Google Cloud Speech API.

4.3 User experiment

To evaluate the system, 19 under graduate or graduate students of varying backgrounds were were asked to used the NDS. 9 subjects used the NDS with our proposed model, and the other 10 used the NDS with the baseline method. Subjects were first shown a video explaining the system, including a demo of its usage. Subjects were then asked to use the system for 5 news topic, which takes about 15 minutes to finish. No instructions on the usage of the system were given. After using the system, subjects were asked to fill a questionnaire.

4.4 Evaluation metric

We compare our end-of-turn prediction model with the baseline implementation by two metric. First is turn-taking speed. We would like to see whether our proposed model can allow the NDS to take turn faster than conventional methods. Time between the end of user voice activity and the beginning of system voice activity can

be used to measure the turn-taking speed. Secondly, we would like to analyze the effect of turn-taking speed on the task of large information transfer though dialog. Our belief is that faster turn-taking will cause the user to be more engaged on the conversation, increasing user feedback. With increased user feedback, the NDS can determine with higher accuracy whether to include or exclude a piece of information. Therefore, we use a measure called the efficiency of information transfer(EoIT), which is a measure that considers how much information the user needed was presented, and how much unnecessary information was excluded by the NDS. EoIT is the harmonic mean of the ratio of sentences $r1$ that were presented and the user considered important, and the ratio of sentences $r2$ that were not presented and the user considered not important.

$$EoIT = \frac{2}{\frac{1}{r_1}+\frac{1}{r_2}}$$

4.5 User experiment results

Fig 1 shows the distribution of turn-taking speed when using our proposed end-of-turn prediction model and when using the baseline method. Results show that using the proposed method, the average turn-taking speed is faster by 782ms. User experience of the length of turn-taking speed also coincide with this result. In the questionnaire, users were asked if they felt the time between the end of their speech and the system's response was long on a 5 point scale, and users who used the NDS with our model gave a lower score with a significant difference ($p < 0.1$). EoIT was calculated for each user. However no significant difference could be shown.

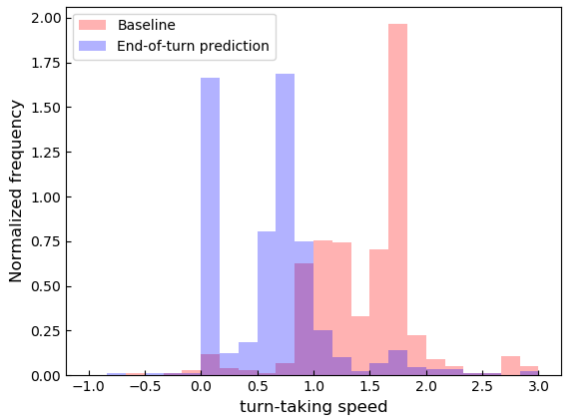

図 1 Turn-taking speed of NDS when using baseline method compared to the proposed end-of-turn prediction model

5. Summary

In this research, we proposed a RNN based model for predicting the end-of-turn of user speech in a spoken dialog system, using prosodoc as well as lexical information as input. By constantly predicting the end-of-turn of the user using the proposed model, a spoken dialog system can take the turn with smaller delay, and interruption of the user compared to conventional turn-taking methods. Through a user experiment on a previously developed spoken dialog system for delivering large amount of information, the proposed model was shown to be able to decrease the turn-taking speed compared to conventional methods. Also, questionnaire showed that users can detect this difference although it is shorter than a second. However despite our belief that a fast paced turn taking should increase the information delivering efficiency, user experiment showed that this is not so. In future research we hope to explore other aspects of turn-taking and its effect on large information transfer.

参考文献

[1] Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K. and Kawahara, T.: Attentive listening system with backchanneling, response generation and flexible turn-taking, No. August, pp. 127–136 (online), DOI: 10.18653/v1/w17-5516 (2018).

[2] Skantze, G.: Towards a General , Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks, No. August, pp. 220–230 (2017).

[3] Nigel Ward, Diego Aguirre, G. C. and Fuentes, O.: Turn-Taking Predictions Across Languages and Genres using an LSTM Recurrent Neural Network, pp. 831–837 (2018).

[4] Ishimoto, Y. and Enomoto, M.: Prosodic Changes Leading to Transition Relevance Place in Spontaneous Utterance, pp. 31–37.

[5] Takatsu, H., Fukuoka, I., Fujie, S., Hayashi, Y. and Kobayashi, T.: A Spoken Dialogue System for Enabling Information Behavior of Various Intention Levels, *Transactions of the Japanese Society for Artificial Intelligence*, Vol. 33, No. 1, pp. DSH–C_1–24 (2018).