Detection of Feeling through Back-channels in Spoken Dialogue

Tatsuya Kawahara[*†], *Masayoshi Toyokura*[*†], *Teruhisa Misu*[†*], *Chiori Hori*[†]

[*]School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[†]National Institute of Information and Communications Technology, Kyoto 619-0288, Japan

We investigate the usage of back-channel information in the information navigation dialogue between an expert guide and a user. By back-channel feedback, we mean the user's verbal short response, which expresses his state of the mind during the dialogue. Its prototypical lexical entries include "*hai*" in Japanese and "yes" or "right" in English, however, we do not count explicit affirmative responses as back-channels.

Previously, there were several works[1, 2] which attempted to automatically generate back-channel responses for smooth communication between the user and the system. Recently, the back-channel information is included in the framework of dialogue act tagging in the game-playing dialogue[3] and meetings[4]. In the information navigation dialogue, in which an expert guide presents a list of recommendation spots, it is expected that the prosodic pattern of the back-channel conveys the para-linguistic information, that is, it suggests the positive/negative feeling on the recommended candidate. We also presume that the human expert guide detects such feelings expressed via back-channels, and chooses to continue the explanation of the current topic if the user seems interested, or change the topic otherwise. Thus, we investigate the back-channel patterns observed in the Kyoto Tour Guide Dialog Corpus.

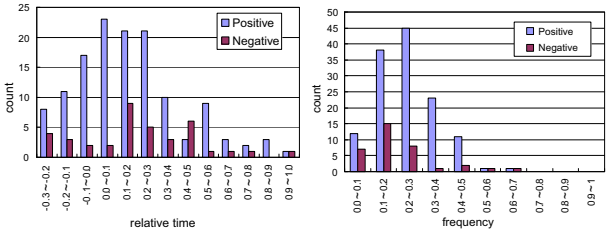
Index Terms: prosody, back-channel, spoken dialogue, feeling detection

1. Episode Segmentation and Tagging

In this dialogue task, the expert guide makes recommendations on the sightseeing spots and restaurants, until the user determines his plan on that day, typically two or three visits and lunch and dinner places. First, we segmented the dialogue session into units called episodes, which correspond to recommendation of one particular spot. The segment starts with a presentation of the spot, followed by detailed explanations along with the users' occasional questions and responses. The episode ends when the user accepts the recommendation, or declines it, or the guide changes the topic. We label a "positive" tag to the episode when the user accepted the spot, which was put into the itinerary of the day; otherwise we label a "negative" tag.

2. Feature Extraction

In this work, we parameterize the back-channel information with two features representing timing and frequency, respectively. Here we assume that the back-channel feedback by the hearer can be inserted around the end of every utterance or IPU (Inter-Pausal Unit) of the current main talker, and measure the time difference of the two events. Note that this value can be negative when the hearer makes a response right before the end of the IPU. The back-channel frequency is defined by the observed count of the responses divided by the number of IPUs so far during the episode segment.



(a) back-channel timing (b) back-channel frequency

Figure 1: Distribution of back-channel timing and frequency

3. Results

We used eight dialogue sessions that have annotation of precise timing information, each lasting about 30 minutes.

The distribution of the relative time of back-channel responses is plotted in Figure 1 (a). We can observe that when a user makes prompt responses, majority of them are positive. For example, if we choose those quicker than 100ms, we could get a precision of 84% (recall: 27%) for positive responses. On the other hand, we hardly get information from the slow responses, since a number of positive responses are made even after 100ms as well as negative responses.

The distribution of the frequency of back-channel responses is shown in Figure 1 (b). It is confirmed that more back-channels suggest positive responses. For example, if we choose those cases more than 0.3, a precision of 86% (recall: 45%) could be obtained for positive responses.

Moreover, we can get a synergetic effect by combining these two features. If we pick up those cases which satisfy either of the above-mentioned conditions, we could get a precision of 84% with a recall of 59% for positive responses. The result demonstrates that we can detect positive responses with high accuracy, and this property is expected to realize a proactive dialogue management which would present detailed information to the current topic without making an explicit confirmation.

4. References

[1] N.Ward. Using prosodic clues to decide when to produce back-channel utterances. In *Proc. ICSLP*, pages 1728–1731, 1996.

[2] N.Kitaoka, M.Takeuchi, R.Nishimura, and S.Nakagawa. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. *J. Japanese Society for Artificial Intelligence*, 20(3):220–228, 2005.

[3] A.Gravano, S.Benus, J.Hirschberg, S.Mitchell, and I.Vovsha. Classification of discourse functions of affirmative words in spoken dialogue. In *Proc. INTERSPEECH*, pages 1613–1616, 2007.

[4] F.Yang, G.Tur, and E.Shriberg. Exploiting dialog act tagging and prosodic information for action item identification. In *Proc. IEEE-ICASSP*, pages 4941–4944, 2008.

Copyright © 2008 ISCA