2ZD-08

Automatic Evaluation of Presenters' Discussion Performance Based on their Heart Rate

Shimeng Peng† Katashi Nagao‡

Department of Media Science, Graduate School of Information Science, Nagoya University†

1 Introduction

Heart Rate Variability (HRV) has emerged enormous potential on cognitive performance evaluation as taking insight into the autonomic nervous system. In this study, we propose that HRV of presenters can be used to effectively evaluate one of the cognitive performance: discussion performance, which consists of several Q&A statements. To validate our opinion, noninvasive device Apple watch was used to collect presenters' HRV, three machine learning models: Logistic Regression, Support Vector Machine, Random Forest has been generated and discussed. Comparative experiments were performed to evaluate discussion performance using traditional semantic data of Q&A statements alone and the combination of HRV and semantic data. We also confirmed the robustness of HRV features on the new dataset.

2 Presenters' Heart-Rate Data Acquisition

2.1 Discussion-mining system

Discussion is usually considered as an effective way for knowledge discovery and information exchange, and is always carried out through participants asking questions and discussion presenters answering them. We call these question and answer pairs as Q&A segments. We think that the more higher-quality answers given by presenters the better discussion performance can be achieved. We previously developed a "Discussion Mining (DM)" system which generates multimedia meeting minutes from recording face-to-face discussion in our lab environment, it includes audio-visual and semantic information of Q&A segments as shown in figure 1 which can be analyzed for the presenters' discussion performance evaluation [1]. Answer-quality of Q&A segments always be evaluated by Natural language processing (NLP) such as analyzing the semantic information of answer statements. However, the personal customaries always result in a low generalization performance of Q&A segments' answer-quality evaluation.

2.2 Heart-rate-data acquisition system

Considering discussion is a kind of cognitive activity which could cause changes in some physiological data, like heart rate variability (HRV) [2]. We developed a presenters' heart-rate (HR) acquisition system based on our DM system to collect the presenters' HR data during their Q&A segments in discussion, HR data would be used to evaluate the answer-quality of Q&A segments later.

A non-invasive device, Apple Watch was employed to acquire presenters' HR data updated in 5-7 sec intervals through the Health Kit framework. We asked presenters to wear on left hand during their discussion, the collected HR information is shown on the Apple Watch's screen as well as synchronously presented on our HR web browser. As shown in Figure 2.

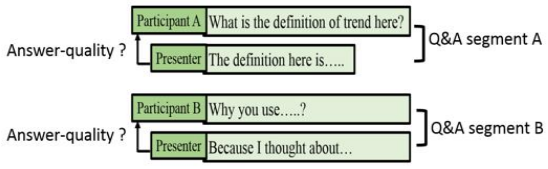

Figure 1: Q&A segments recorded by DM system

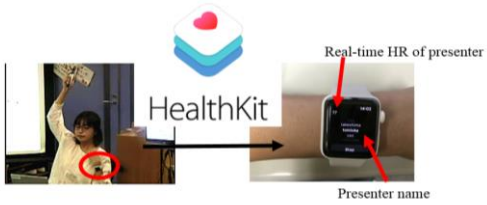

Figure 2: Presenters' heart-rate (HR) acquisition

3 Evaluation Experiments

3.1 Experimental data

We totally collected two times of sample data. At first, we collected 9 presenters' discussion data from 9 lab-seminar discussion. 117 Q&A segments were extracted and then the answer-quality of them were evaluated by the participants who asked the questions by gave a score based on a five-point scale: 5= very good, 1= very bad. There were 51 low-quality answers with scores from 1-3, and 66 high-quality answers with scores from 4-5.

18 HRV features were computed from the complete Q&A periods as well as the question and answer periods separately, which include mean, standard deviation (std.), and root mean square successive difference (RMSSD) from all of the three periods as well as each periods' HRV upward or downward trends decided by calculating the difference number between two adjacent HR points. We also divided the HR data into 9 ranges: less than 60bpm, 60-70bpm… and more than 130bpm, the mean and std. were also

発表者の心拍数を用いた議論の達成度の自動評価
†彭詩朦　長尾確
†名古屋大学　大学院情報科学研究科

calculated in each range to describe the HR appearance-frequency distribution.

3.2 Evaluation experiments and results

We generated three kinds of binary classification models: Logistic regression (LR), Support vector machine (SVM) and Random forest (RF) as the evaluation models. 80% of Q&A segments were randomly selected as the training dataset and the remaining one used as the test dataset.

Considering that using all of the HRV features could decrease the performance of evaluation models, we recurred to RFE and the RFECV methods to decide the most suitable feature subset for each models which will resulted in a highest F-measure. There were 7 HRV features selected: All mean, Answer trend, All RMSSD, Freq answer std., Answer std., Question trend, and All trend, which exhibited the largest effect on all three models.

Taking insight into our evaluation results, we obtained a 0.79 F-measure for the LR model, a 0.8053 F-measure for the SVM model and a 0.87 F-measure for RF model. Considering all of the three evaluation models, the HRV data of presenters showed an outstanding evaluation performance of Q&A segments' answer-quality, especially the RF model which achieved a 0.87 F-measure and has exhibited a superior evaluation performance.

This results provide us evidence that HRV data of presenters can be used to effectively evaluate the answer-quality of Q&A segments and as a discussion performance evaluation method.

4 Comparative Experiments

4.1 Comparative experiments on semantic feature alone and the combination of these two types of data

In order to further argue that whether the HRV features of presenters show better discrimination performance regarding the Q&A segments' answer-quality evaluation than semantic features extracted from presenters' statements, we conducted two comparative experiments by generating LR, SVM, RF models based on the semantic features of Q&A statements alone and the combination of them

We took advantage of 1246 Q&A segments data recorded before in our lab environment and evaluated the answer-quality with the same method as section 3, 993 high-quality and 253 low-quality Q&A segments were obtained. Morpheme bigram was generated based on these survey Q&A segments and several bigrams were extracted as the certain semantic features if their occurrences were much higher than 0.15%, there were 14 semantic features selected for evaluating the Q&A segments' answer-quality.

The results are shown as table 1, all of the models received low F-measure when used semantic features alone and RF model obtained a relative higher F-measure even though only 0.583. However, we were surprised find that combined these two kinds of data increased the evaluation performance of Q&A segments' answer-quality in a certain extent.

Table 1: Evaluation-performance comparison of HRV and semantic features and their combination for each evaluation model

model	F-measure		
	HRV features	Semantic features	Combination of HRV and semantic features
LR	0.79	0.5	0.79
SVM	0.8053	0.54	0.833
RF	0.87	0.583	0.916

4.2 HRV data robustness evaluation experiments

To verify the robustness of HRV features, we collected another 8 times' presentation data and extracted a total of 66 Q&A segments. 20% Q&A segments were randomly selected from the new dataset and the previous test dataset together to become new test dataset, remaining Q&A segments with the previous training dataset were used as new training dataset. We generated new evaluation models of LR, SVM, and RF by using the 7 meaningful HRV features, we got a 0.756 F-measure of LR model, a 0.81 F-measure of SVM model and a 0.837 F-measure of RF model. From the evaluation results, HRV data of presenters still show a well discrimination performance on new dataset even though has a slight decline of F-measure. We take insight into the Recall score on low answer- quality reorganization in these three HRV models, we achieved a 0.56 Recall of LR model, a 0.78 Recall of SVM model, a 0.65 Recall of RF model, which provides us with favorable evidence that HRV features can be used to effectively recognize the low answer-quality and as a discussion performance evaluation method.

We also verified the robustness of semantic features on this new dataset. We got a similar F-measure with the results on old dataset, however, we got a really low Recall scores, such as 0.06 Recall of LR model, 0.24 Recall of SVM model and 0.48 Recall of RF model, which indicated that the semantic features can not recognize the low quality Q&A segments.

5 Conclusion

In this study, we firstly developed a presenters' HR data acquisition system. Then, we conducted experimental investigations to validate that HR data of presenters can be used to effectively evaluate the answer-quality of Q&A segments and as a discussion performance evaluation method with robustness.

Reference

[1] K. Nagao, K. Inoue, N. Morita and S. Matsubara. Automatic Extraction of Task Statements from Structured Meeting Content. Proc. of the 7th International Conference on Knowledge Discovery and Information Retrieval, pp.307-315, 2015.

[2] Anderson, K. P. (1995). Vagal control of the heart: Experimental basis and clinical implications. Futura Publishing Company.