Predicting Emotional Responses From Spontaneous Social-Affective Interaction Data *

☆ Nurul Lubis, Sakriani Sakti, Graham Neubig, Koichiro Yoshino, Satoshi Nakamura (NAIST)

1 Introduction

The study of social-affective communication is concerned with the role of emotion in human interaction. Emotion plays a two way role in human conversation; we express our emotion and are affected by our conversational partner. This is yet to be completely replicated in human-computer interaction (HCI).

In addition to the more traditional works on recognition and simulation, there has recently been an increasing interest in emotional triggers, studying what causes emotion in the first place. A recent study by Hasegawa et al. [1] addresses this issue by predicting and eliciting emotion in online conversation. We have followed up this study by performing a similar task on natural spoken conversation. We recognized emotion based on a speaker's utterances and analyzed the conversational partner's as the emotional triggers [2, 3].

In this paper, we extend upon the previous works by trying to predict a person's emotional reaction in a social-affective conversation. We examine not only the person having the reaction, but also their conversational partner. The prediction is aimed to accommodate the ability to predict the response to an emotion trigger. Two languages are examined: English and Indonesian.

2 Task Description

Figure 1 presents the workflow of this paper. We extract tri-turns from spontaneous conversation data and try to predict the emotional response occurred within.

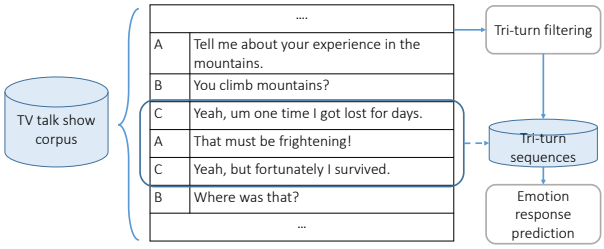

Fig. 1 *Overview*

We collect the data from television talk shows, which contain real conversation with natural emotion occurrences. We utilize the English-Indonesian TV Talk show corpus [4], containing 62 minutes of English interaction and 94 minutes of Indonesian.

In natural conversation, to properly analyze the fluctuation of emotion, it is necessary to ensure that the observed sequences of conversation are in response to each other. Thus, we group consecutive sequences of conversation into a unit called a tri-turn [5]. Three consecutive sequences of speech in a conversation is considered a tri-turn when the second sequence is in response to the first, and the third is in response to the second. On each tri-turn, we observe the first two turns and predict emotion response occurring on the third.

To observe the data from the dialogue perspective, we define a set of dialogue acts adapted from [6] to describe the structure of discourse. The 17 dialogue act labels are given in Table 1.

Table 1 *Dialogue acts*

id	Dialogue Act	id	Dialogue Act
stat	Statement	rept	Repeat Phrase
opi	Opinion	ack	Acknowledgement
back	Backchannel	thnk	Thanking
Qyno	Yes-No Question	apcr	Appreciation
Qopn	Open Question	aplg	Apology
Qwh	Wh Question	hdg	Hedge
Qbck	Backchannel Question	drct	Directive
conf	Agree/Confirm	abdn	Abandoned
deny	Disagree/Deny		

We define the emotion scope based on the circumplex model of affect [7]. Two dimensions of emotion are defined: valence and arousal. Valence measures the positivity or negativity of emotion. On the other hand, arousal measures the activity of emotion. From the valence-arousal scale, we derive 5 common emotion terms: happiness, anger, sadness, contentment, and neutral. We utilize both points of view of emotion to label the data: the value of valence and arousal; and the degree of presence of the emotion classes.

The emotional responses considered in this paper can be categorized into three events. As the events can happen simultaneously in different observations, we examine the emotion class, valence

* 自然生起する社会的・感情的インタラクションデータからの感情反応の予測

level, and arousal level separately. For each observation, we try to predict whether it stays *constant*, changes from postive to negative (*drop*), or changes from negative to positive (*raise*).

3 Automatic Prediction of Emotional Responses

As mentioned previously, on a tri-turn, we observe the first two turns to predict emotion response occurring on the third. This means, on each tri-turn, we stack the features of the first two turns to gather information of the context. The features include acoustic features and dialogue act labels. To gather information from the speech, we extract acoustic features as defined in the INTERSPEECH 2009 emotion challenge [8] using the openSMILE feature extractor.

To balance the number of instances and features, we perform correlation-based feature extraction and linear discriminant analysis of our feature set. After reducing the dimensions, we train a neural network classifier with one hidden layer using Theano and the PDNN toolkit.

To summarize, for all tri-turns, based on the acoustic features and dialogue act labels of the first two turns, we attempt to predict whether emotion events defined occurred on the third. With three classes for prediction, we have a chance rate of 33.33%.

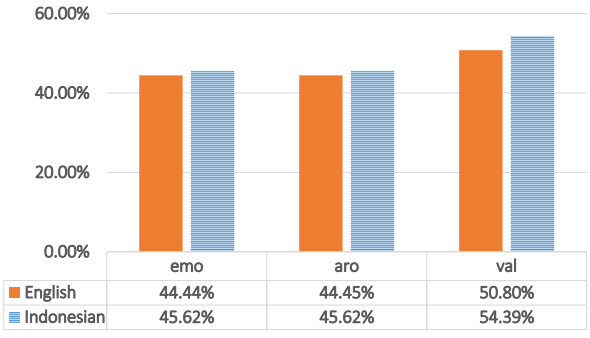

Fig. 2 *Performance of response prediction*

Figure 2 presents our response prediction accuracy. For both languages, the prediction regarding valence has the highest performance, followed by arousal and emotion class. This could mean that valence and its fluctuations are expressed in speech more than is the case for arousal.

The suboptimal performance is likely due to the limited amount of data used in this study. Inherently, there are numerous factors that leads to a change of emotion in a conversation. This means that to properly recognize patterns for such events, a large number of features are required. It is likely that if we could prepare more data in the future, the accuracy will increase significantly.

4 Conclusion

In this paper, we presented a study on social-affective communication for automatic prediction of emotional responses. In providing users the most natural HCI, emotion is an aspect that should not be overlooked. Our experiment on automatic prediction offers an approach in equipping conversational agents and dialogue systems with social-affective awareness.

In future studies, we hope to include more modalities of interaction in observing the dynamics of emotion in interaction, such as textual and visual features. We also hope to incorporate this information directly into a dialogue system.

Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 24240032 and 26870371, as well as by a joint research project with Yanmar Co., Ltd.

References

[1] Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. Predicting and eliciting addressee's emotion in online dialogue. In *ACL (1)*, pages 964–972, 2013.

[2] Nurul Lubis, Dessi Lestari, Sakriani Sakti, Graham Neubig, Tomoki Toda, Ayu Purwarianti, and Satoshi Nakamura. Emotion and its triggers in human spoken dialogue: Recognition and analysis. *Proc IWSDS*, 2014.

[3] Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. A study of social-affective communication: Automatic prediction of emotion triggers and responses in television talk shows. In *Proc ASRU*, 2015.

[4] Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Construction and analysis of social-affective interaction corpus in english and indonesian. In *Proc Oriental COCOSDA*, 2015.

[5] Nio Lasguido, Sakriani Sakti, Graham Neubig, TODA Tomoki, and Satoshi Nakamura. Utilizing human-to-human conversation examples for a multi domain chat-oriented dialog system. *IEICE TRANSACTIONS on Information and Systems*, 97(6):1497–1505, 2014.

[6] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[7] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[8] Björn Schuller, Stefan Steidl, and Anton Batliner. The INTERSPEECH 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315, 2009.