原著論文

Original Paper

Emotional Triggers and Responses in Spontaneous Affective Interaction: Recognition, Prediction, and Analysis

Nurul Lubis — Nara Institute of Science and Technology
nurul.lubis.na4@is.naist.jp

Sakriani Sakti — (affiliation as previous author)
ssakti@is.naist.jp

Koichiro Yoshino — Nara Institute of Science and Technology, Japan Science and Technology Agency
koichiro@is.naist.jp

Satoshi Nakamura — Nara Institute of Science and Technology
s-nakamura@is.naist.jp

keywords: emotion recognition, emotion trigger, affective computing, dialogue system

Summary

To completely mimic the naturalness of human interaction in Human-Computer Interaction (HCI), emotion is an essential aspect that should not be overlooked. Emotion allows for a rich and meaningful human interaction. In communicating, not only we express our emotional state, but we are also affected by our conversational counterpart. However, existing works have largely focused only on occurrences of emotion through recognition and simulation. The relationship between an utterance of a speaker and the resulting emotional response that it triggers is not yet closely examined. Observation and incorporation of the underlying process that causes change of emotion can provide useful information for dialogue systems in making a more emotionally intelligent decision, such as being able to take proper action with regard to user's emotion, and to be aware of the emotional implication of their response. To bridge this gap, in this paper, we tackle three main tasks: 1) recognition of emotional states, 2) analysis of social-affective events in spontaneous conversational data, to capture the relationship between actions taken in discourse and the emotional response that follows, and 3) prediction of emotional triggers and responses in a conversational context. The proposed study differs from existing works in that it focuses on the change of emotion (*emotional response*) and its cause (*emotional triggers*) on top of the occurrence of emotion itself. The analysis and experimental results are reported in detail in this paper, showing promising initial results for future works and development.

1. Introduction

The *appraisal theory of emotion* argues that most of our emotional experiences are the result of a cognitive process, unconscious or controlled, of evaluating situations and events [Ellsworth 03, Scherer 01]. In social situations, this process occurs continuously between conversational partners, creating a dynamic loop of expressions of emotion, which triggers a change in emotional state as a response, and so forth.

It is argued that humans also impose this emotional aspect in their interaction with computers and machines [Reeves 96]. They treat them politely, laugh with them, and sometimes get angry or frustrated at them. Many works and studies have attempted to equip computers with emotional capabilities to reciprocate with humans in this regard. In Human-Computer Interaction (HCI), in particular for dialogue systems and conversational agents, this topic continues to gain traction.

The most widely researched sub-areas of social-affective communication are *emotion recognition* and *emotion simulation*. *Emotion recognition* allows a system to discern the user's emotions and address them in giving a response [Han 15, Tielman 14]. On the other hand, *emotion simulation* helps convey non-verbal aspects to the user for a more believable and human-like interaction, for example to show empathy [Higashinaka 08] or personality [Egges 04].

In addition to the more traditional works on emotion recognition and simulation, there has recently been an increasing interest in *emotion elicitation*, or *emotional triggers*, in which a computer system attempts to trigger a certain emotion from the user through the interaction. Previous work by [Hasegawa 13] reports proper elicitation of basic emotions by training individual models that "translates" user's input into the eliciting response. On the other hand, Skowron et al. have studied the impact of different affective personalities in a text-based dialogue system [Skowron 13]. They reported consistent impacts with the corresponding personality in humans.

However, the existing studies have not yet observed the relationship between an utterance of the speaker, which acts as a stimulus evaluated during appraisal, and the resulting emotion by the end of appraisal. Knowledge of this process can provide useful information for a system to be more emotionally intelligent in communicating with the user, for example, to refrain from provocative responses and to seek pleasing ones.

In this paper, we extend previous studies by analyzing and predicting emotional events in spontaneous social-affective conversation. We approach this problem by examining the relationship between emotional triggers and responses in a spontaneous conversational data; we observe fluctuations of emotion and their correlation to actions taken in discourse. The proposed study differs from existing works in that it focuses on the change of emotion (*emotional response*) and its cause (*emotional triggers*) on top of the occurrence of emotion itself.

In particular, we tackle three main tasks:

(1) Recognition of emotional states, with which an emotional response can be measured.

(2) Analysis of social-affective events in spontaneous conversational data to capture the relationships between dialogue actions and emotional events.

(3) Prediction of emotional triggers and responses given a conversational context. This task could provide dialogue systems with the ability to take proper action with regard to user's emotion and the emotional implication of their response.

Prior to the experiments, we construct a spontaneous social-affective corpus to represent natural emotion-rich human interaction. We collect recordings of television talk shows in English and carefully annotate them with relevant information needed in the proposed study. The resulting corpus, experiments, and analysis should serve to minimize the gap between emotional aspects in human communication and HCI.

2. Data Construction

Although there has been an increase of interest in constructing corpora containing social interactions [Douglas-Cowie 07, Ringeval 13], there is still a lack of spontaneous and emotionally rich corpora. To bridge this gap, we construct a corpus of spontaneous social-affective interaction in the wild. We utilize various television talk shows containing natural conversations and real emotion occurrences. Interactions in talk show settings represent typical social conversation, where a small number of speakers are involved and various emotion-inducing topics are discussed.

From this data, we would like to observe three major aspects of the conversation: 1) emotional states of the speakers, 2) the emotional events, how emotion fluctuates, affected by and affecting the conversation, and 3) the social aspects, concerning the conversational actions and reactions between the speakers. We annotate the data with relevant information to allow these observations. This chapter elaborates the steps in constructing the corpus, starting from data collection and concluding with annotation.

2·1 Data Collection

We aim to capture realistic emotion that is applicable to human-computer interaction. To achieve this, we construct our corpus from a number of talk shows covering various topics of discussion, such as life experiences and politics. This approach allows the collection of data that contains natural conversations and real emotion occurrences without the cost of recording and recruitment of participants.

We collect three episodes of two of the most popular American television talk shows. One of the shows focuses on life experiences of celebrities and their struggle with weight loss in the public eye. The other two contain dialogues about overcoming severe family problems, one involving pathological lies and one dealing with physical and emotional problems. All of the these topics are inherently emotional for the speakers involved. These talks demonstrate real problem identification, experience sharing, negotiation, and advice offers. As the talk shows involve more than one speaker, they contain many exchanges of both positive and negative emotional expressions.

In total there are 12 speakers, consisting of 4 male speakers and 8 female. The collected data amounts to 1 hour, 2 minutes, and 19 seconds of speech. Video recordings of the show are obtained, but stripped down to audio only as we are currently focusing on speech data. Audio is available at 16 kHz and 16 bits per sample.

2·2 Annotation Procedure

We impose rigorous quality control in data annotation to ensure the result's consistency. Most importantly, we bear in mind that language and culture affect how emotion is perceived and expressed in an interaction. We carefully select 3 annotators, each is required to be 1) a native speaker of the language used in the show, i.e, English, and 2) knowledgeable of the culture appearing in the talk shows. With these requirements, we try to ensure that the annotators can observe emotion dynamics of the interaction to the furthest extent. To ensure consistency, we have each annotator work on the full corpus. Figure 1 gives an overview of the annotation procedure.

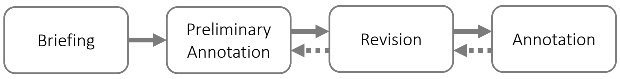


Fig. 1 Overview of annotation procedure. Dashed lines denotes possible iterative loop.

Before annotating the corpus, the annotators are given a document explaining the task and its goal in detail. The document provides theoretical descriptions of emotion and dialogue acts, as well as a number of examples. Afterwards, they perform preliminary annotation on a small subset of the corpus. The reason for this is twofold: this step allows the annotators to get used to the task, and by verifying the preliminary result's quality and consistency, we will be able to confirm whether the annotators have fully understood the guidelines.

We manually screen the preliminary result and give feedback to the annotators accordingly. They are asked to revise inconsistencies with the guidelines if there are any. Once the results are verified, the annotators then work on the rest of the corpus. We perform the screen-and-revise process repeatedly to achieve a tenable result.

2·3 Annotation Labels

§ 1 Emotion

Defining and structuring emotion is essential in observing and analyzing its occurrence in a conversation. We define the emotion scope based on the circumplex model of affect [Russell 80], depicted in Figure 2. Two dimensions of emotion are defined: *valence* and *arousal*. Valence measures the positivity or negativity of emotion; e.g., the feeling of joy is indicated by positive valence while fear is negative. On the other hand, arousal measures the activity of emotion; e.g., depression is low in arousal (passive), while rage is high (active).

This model describes the perceived form of emotion intuitively. Furthermore, it is easily adaptable and extendable to either discrete or other dimensional emotion definitions. The long established dimension are core to many works in affective computing, such as [Trigeorgis 16, Zhao 16].

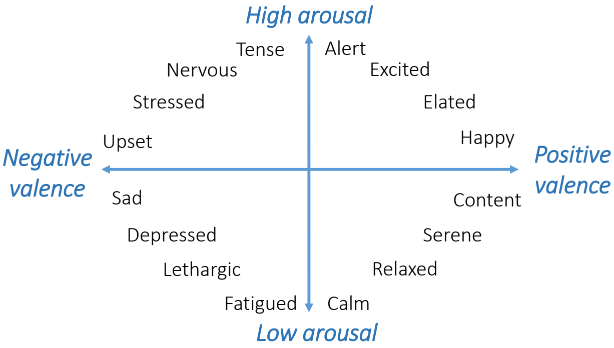


Fig. 2 Emotion classes and dimensions.

Following this model of emotion, the emotion annotation of the corpus consists of the level of valence (`val`) and arousal (`aro`). The value of each dimension can be as low as -3 (most negative) and as high as 3 (most positive) with a discrete step of 1. This balances the resolution of the emotion information with the cognitive load of the annotators. We perform the annotation at sentence level. In this work, we define a sentence in a dialogue as a continuous utterance of a speaker until 1) a full sentence is produced, or 2) it is proceeded by a different speaker.

§ 2 Dialogue Act

To analyze the social aspect, it is necessary to observe the relationship between the utterances in a dialogue. To do this, we define a set of dialogue acts adapted from [Stolcke 00] to describe the structure of discourse. To avoid sparsity, we reduce the original set of dialogue acts from 42 to 17 by grouping together similar acts, such as Yes-No-Question and Declarative Yes-No-Question. The 17 dialogue act labels are given in Table 1. As with the emotion annotation, the dialogue act labeling is performed at sentence level.

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		

2·4 Annotators Agreement

To analyze the reliability, we measure an agreement metric for each annotation label set. We use Pearson's correlation for the numeric emotion labels, and Fleiss' Kappa for the nominal dialogue act labels. Numeric variables represent measurable quantity, whereas nominal variables represent categories that have no logical order.

§ 1 Emotion Labels

We calculate mean Pearson's correlation coefficients r of the three annotators. Pearson's r measures the strength and direction of linear relationship between two variables. An absolute value between 0.0 and 0.3 is interpreted as weak correlation, greater than 0.3 up to 0.5 as moderate correlation, and higher than 0.5 as strong correlation. We observe a moderate correlation for arousal annotation (0.32) and a strong correlation for valence (0.64) in the result of emotion annotation.

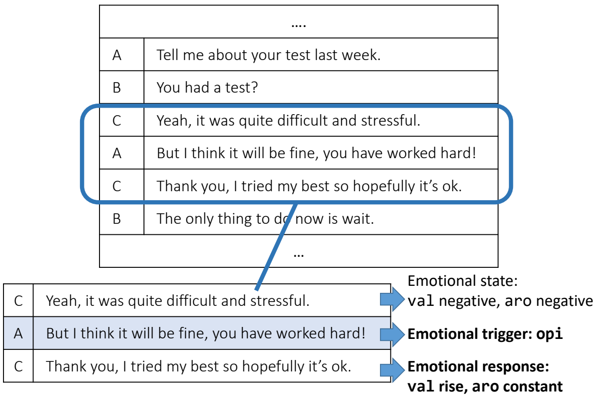
§ 2 Dialogue Act Labels

We calculate Fleiss' kappa κ of the three annotators. κ measures the inter-annotator agreement of nominal variables when more than two annotators are employed. The dialogue act annotation of the corpus results in a κ of 0.54. According to interpretation of κ [Fleiss 71], this annotation is considered to have a moderate agreement.

2 · 5 *Social-Affective Events*

Natural conversation can be unordered and disconnected from one turn to the next. Thus, to properly analyze the fluctuation of emotion in a conversation, it is necessary to ensure that the observed sequence of dialogue turns are in response to each other.

We group three dialogue turns that are in response to each other into a unit called a *triturn* [Lasguido 14]. That is, within three consecutive dialogue turns, the second turn is in response to the first, and the third is in response to the second. We ensure that there are exactly two speakers in a triturn, forming an A-B-A dialogue speaker format. In this study, the triturn format is useful in observing emotional triggers and responses in a dialogue.



	....
A	Tell me about your test last week.
B	You had a test?
C	Yeah, it was quite difficult and stressful.
A	But I think it will be fine, you have worked hard!
C	Thank you, I tried my best so hopefully it's ok.
B	The only thing to do now is wait.
	...

C	Yeah, it was quite difficult and stressful.	Emotional state: val negative, aro negative
A	But I think it will be fine, you have worked hard!	**Emotional trigger: opi**
C	Thank you, I tried my best so hopefully it's ok.	**Emotional response: val rise, aro constant**

Fig. 3 An example of a triturn and its social-affective events, marked in bold, extracted from a conversation.

Figure 3 illustrates a triturn example and the social-affective events that are observable within. We refer to the change of emotion from the first turn to the third as *emotional response*. As emotional responses, we separately examine valence and arousal responses since they can occur at the same time. We categorize changes of each emotion dimension into three events: rise, drop, or constant. On the other hand, we consider the second turn to be the *emotional trigger* of the response. As emotional triggers, we consider the set of 17 dialogue acts in Table 1.

To form the triturns from the sentence-level segments, we first concatenate consecutive sentences of the same speaker until a change of speaker occurs. The concatenated sentences form the dialogue-turn segments. Afterwards, we form triturn units from dialogue turns with A-B-A pattern. From the data, 626 triturns are collected.

3. Speech Emotion Recognition

Emotion recognition is a fundamental part of an affective dialogue system. Without the ability to track user's emotional state, incorporation of emotion could instead degrade the user experience and result in erroneous actions of the system. Furthermore, the ability to identify emotional states is the pre-requisite for identifying changes of emotion over time. In this chapter, we present an experiment on recognizing affective states from speech as the first step in studying social-affective aspects in human communication.

At this stage of our research, we limit our scope to the speech channel, which has been argued to be one of the most salient channels to communicate emotion [Schuller 03]. It is also the fastest and most natural way of communication between humans [El Ayadi 11]. It communicates a wide array of information, and thus is suitable for representing emotional information in spontaneous speech. Furthermore, speech features can be reliably and automatically computed for any speech data, unlike human annotated information, such as transcriptions, which requires extensive labor and time to obtain. The result and analysis from this experiment may provide potential information to aid the emotional triggers and responses task.

3 · 1 *Experimental Set Up*

We perform the emotion recognition task at sentence level. We simplify the emotion recognition problem by discretizing the affective dimensions values into three classes: positive, neutral, and negative. The reason is twofold: 1) To capture the global value of the utterance, and 2) to avoid class sparsity in the data. This simplification is a trade-off decision between the resolution of recognized emotions and the model complexity. We believe that incremental steps in tackling a problem will result in a better recognizer in the long run, thus we decide on the less complex problem at this stage of the research.

We separate emotion recognition into two tasks: valence and arousal recognition. We map the numeric annotations to the three classes. The positive class corresponds to the positive-valued annotation (3 to 1), the neutral class to the zero-valued (0), and negative class to the negative-valued (-1 to -3). Subsequently, we perform majority voting between the annotators. Segments with three different votes are excluded to avoid potentially ambiguous emotion occurrences. In total, we obtain 1158 segments for the valence recognition task, and 1105 for arousal. We divide the total with a 80:10:10 ratio into training, development, and test sets.

We extract global features of each utterance using the openSMILE toolkit [Eyben 10]. As classification features, We extract two speech feature sets widely used in emotion recognition tasks and compare the results: INTERSPEECH 2009 baseline features (IS09) [Schuller 09a] and the extended Geneva Minimalistic Acoustic Parameter Set (eGemaps) feature sets [Eyben 16].

The IS09 feature set is described in Table 2. It includes the most common yet promising feature types and functionals covering prosodic, spectral, and voice quality features. On the other hand, the eGemaps feature set is proposed as a reduced acoustic feature set, containing only knowledge-based selected features that are highly potential in indexing affective signals, and proven to be effective in previous studies [Eyben 16]. This feature set includes parameters related to frequency (pitch, jitter, formant), energy (shimmer, loudness, HNR), and spectral information (alpha ratio, Hammarberg index, MFCC 1-4, spectral scope, format relative energy and bandwidth, spectral flux).

Table 2 Baseline feature set of the INTERSPEECH 2009 emotion challenge [Schuller 09a].

LLD (16 · 2)	Functionals (12)
(△) ZCR	mean
(△) RMS Energy	standard deviation
(△) F0	kurtosis, skewness
(△) HNR	extremes: value, rel. position, range
(△) MFCC 1-12	linear regression: offset, slope, MSE

We train an SVM classifier with radial basis function (RBF) kernel using each feature set using the libSVM library [Chang 11]. Prior to training, we scale the features into a $\{0, 1\}$ range to avoid overpowering of features that have a higher value range. Furthermore, we perform parameter optimization with grid search to find the optimal value of C, the cost of misclassification, and γ, the free parameter of the Gaussian RBF kernel. These steps are recommended in [Hsu 03] and have been shown to be effective in SVM classification experiments.

The data distribution of the test set is as follows. For the arousal task, the test set consists of 11 negative, 22 neutral, and 78 positive data points. For the valence task, the test set consists of 74 negative, 23 neutral, and 19 positive data points. As baseline chance level, we consider a model that always predicts the class with highest probability in the data, i.e., the largest class. We show the model performance through its recognition precision, recall, and F1 score on the test set, and comparing it to the baseline chance level. We use weighted averaging based on class support to obtain the final measurements of the experiment. This allows us to take into account the different size of the classes.

3·2 Experimental Results

The optimized SVM parameters of the models are summarized in Table 3. The models trained on the IS09 feature set tend to have larger values of C and smaller γ compared to that of the eGemaps model.

Table 3 Optimized parameters of the models.

Emotion dimension	Feature	C	γ
Arousal	IS09	2^5	2^{-13}
	eGemaps	2^3	2^{-5}
Valence	IS09	2	2^{-5}
	eGemaps	2	2^{-3}

Table 4 presents the performance of emotion recognition. Interestingly, regardless of the much bigger number of features, the IS09 set outperforms the eGemaps set. This suggests that some features that are present in IS09 but absent in eGemaps are helpful in recognizing emotion.

Table 4 Precision, recall, and F1 scores of emotion recognition in percent (%). Best F1 score on each task is boldfaced.

Emotion dimension	Model		Prec.	Rec.	F1
Arousal	Baseline		49.3	70.2	58.0
	Proposed	IS09	72.8	76.5	**72.0**
		eGemaps	63.5	73.8	67.6
Valence	Baseline		40.6	63.7	**49.6**
	Proposed	IS09	49.0	37.9	40.9
		eGemaps	45.3	34.4	38.5

For the arousal task, the recall appears to be higher than precision, while the opposite is true for the valence task. On the arousal prediction, our proposed model successfully surpasses the baseline chance level with a wide margin. However, the arousal recognizer is lower by an absolute 8.7%. This result agrees with previous works on speech emotion recognition which reported noticeably lower recognition performance for valence compared to arousal [Trigeorgis 16]. This signals that valence is characterized more by other communication clues, for example facial expression or content words, while arousal is directly expressed by alternating the speech prosody. For example, humans tend to lower their speech volume to express deactivated emotions and louder to express activated ones.

A close benchmark we could find is [Schuller 09b], where similar acoustic features are used in combination with an SVM classifier with polynomial kernel and sequential minimal optimization. However the task slightly differs. Arousal and valence are viewed as a binary classification problem: low or high for arousal, positive or negative for valence. The binary classification accuracies on the spontaneous English speech corpus are 55.0% for arousal and 49.9% for valence. However, as noted in [Zeng 09], the differences of experimental conditions should be considered when comparing emotion recognition performances.

Readers are referred to other literatures, such as [Anagnostopoulos 15, El Ayadi 11], for a comprehensive survey and comparison on emotion recognition works. Each research effort is aimed to capture emotion in a specific set of circumstances, creating differences between the used corpora as well as the experiments. Furthermore, different emotion models result in different annotation schemes (e.g., dimensional traces or emotion classes), leading to distinct recognition problems (e.g., classification or regression). Each of these experimental designs should be taken into account in comparing model performances.

4. Analysis of Emotion Dynamics

In the analysis, we compute statistical relationships between emotional responses and triggers. We investigate: 1) whether there are dialogue actions that heavily characterize any emotional response, and 2) whether the emotion of the trigger is correlated with the emotional response.

4·1 *Emotional Triggers*

First, we take a look at the distribution of all emotional triggers, regardless of the emotion event they elicit. Figure 4 visualizes the percentage of dialogue act frequencies in the data, sorted from highest to lowest. Statement has the highest frequency, followed by backchannel, confirmation, and yes-or-no question.

The high frequency for backchannels and questions show that such actions provide a way to emotionally engage with the counterpart. Furthermore, providing new information through statements in conversation can also elicit an emotional response from the counterpart. Table 5 includes an example taken from the corpus which demonstrates this finding: the levels of valence and arousal are increasing for both the host and the guest after backchannel and question following up a sentence.

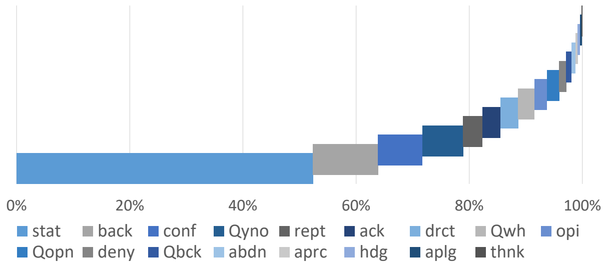


Fig. 4 Dialogue act triggers of all emotion events, ordered by frequency.

Table 5 Emotional responses in a conversation taken from the corpus. Notice the rising of valence and arousal, showing emotional engagement.

Speaker	Transcription	act	aro	val
Guest	Well I still have a lot of clothes in my closet I really shouldn't have	stat	0	-1
Host	Yeah	back	-1	0
Guest	But—yeah	conf	1	1
Host	Why?	Qwh	0	1
Guest	Just [inaudible] I want to say just in case but I don't think so 'cause I really think I got it conquered this time	opi	2	2

Second, we look into the details of each emotion event. To examine the connection between dialogue acts and emotional events, we adapt the term frequency-inverse document frequency (TF-IDF) formula as shown in Equation (1) to measure the importance of each dialogue act in triggering a certain emotion event. The adapted formula is written as

$$\mathrm{tfidf}(d,t,T) = f(d,t) \times \log \frac{\{t \in T\}}{1 + \{t \in T : d \in t\}}, \quad (1)$$

where d is the dialogue act, t is the emotion event, and T is the collection of events. $f(d,t)$ denotes the raw frequency of d in t. This score is calculated for each dialogue act on each emotion event. This score can inform us if a dialogue act characterizes a particular emotion event. As the frequency of stat is high for all types of change, we separate statements according to the speaker's valence (i.e., positive, negative, and neutral), and calculate their scores accordingly.

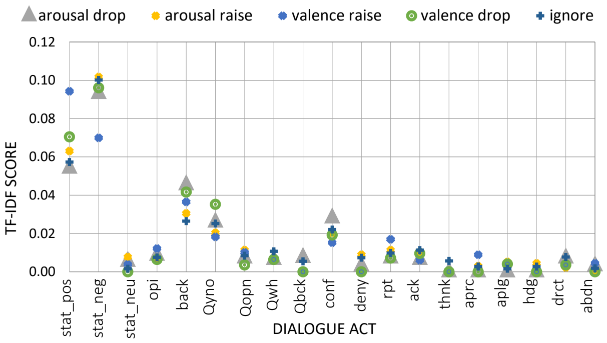


Fig. 5 Dialogue acts scores for all emotion events.

Figure 5 visualizes the TF-IDF scores, where a point in the graph corresponds to the TF-IDF score for a dialogue act on an emotion event. Overlapping dots for a dialogue act mean the TF-IDF score is similar for all emotion events. When this is observed, we can conclude that the dialogue act does not characterize a particular emotion event.

Given the constructed corpus, we mostly observe overlapping points within a dialogue act, except for those with high frequency in the corpus (e.g. `stat`, `back`, and `conf`). However, we think it is possible that the lack of characterization for the majority of the triggers are due to their scarcity in the collected data. Additional data could potentially transform the graph to better reflect the true relationship between dialogue acts and emotional responses. Currently, we can observe that positive-valenced sentences are salient in the events of raised valence, while negative-valenced ones are salient in raised arousal.

4·2 *Emotional Response*

In this section, we investigate whether the emotional content of the trigger is correlated with the corresponding emotional response. As statements make up the largest part of the collected triggers, we take a closer look at them by first grouping them according to the speaker's emotion. Afterwards, we plot them according to the emotional response they trigger.

This distribution is shown in Figure 6. The changes of valence and arousal are shown on the x- and y-axes, respectively. Statements with positive emotion (circles) spread to the upper right and lower left quadrants, showing that they tend to cause an increase or decrease of both valence and arousal at the same time. On the other hand, statements with negative emotion (crosses) tend to give the opposite effects to valence and arousal, i.e., causing an increase of arousal together with a decrease in valence, or a decrease in arousal together with an increase in valence.

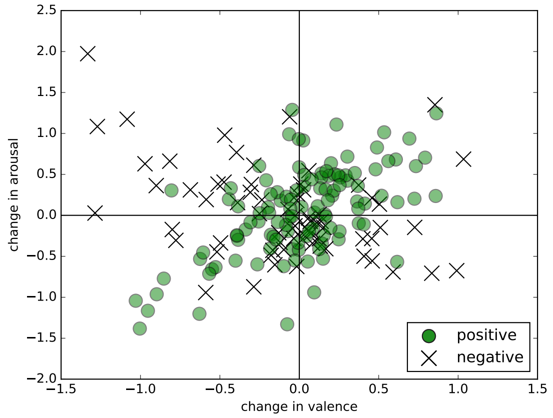


Fig. 6 Emotion of statements with respect to the emotional response it triggers.

Table 6 Average and variance of emotional responses for sentences with positive and negative emotions.

Sentence type	Emotional response	Average	Variance
Positive emotion	Valence	0.014	0.153
	Arousal	0.049	0.295
Negative emotion	Valence	-0.100	0.261
	Arousal	0.040	0.298

This tendency is affirmed by Pearson's correlation coefficient r between the changes in valence and arousal for both types of statements: 0.59 for positive, and -0.36 for negative statements.

We also investigate the distributions of all emotional responses, summarized in Table 6. On average, the emotional responses are rather weak for any type of sentence and emotional dimension, with values that are close to zero, except that of arousal change caused by sentences with negative emotion, which has an average of -0.100. We also observe higher variance for changes in arousal compared to that of valence.

5. Automatic Prediction of Social-Affective Events

In this task, we attempt to model the emotional events in a conversation. The ability to predict this pattern can provide useful information for dialogue systems in making a more emotionally intelligent decision, by being aware of the emotional implication of their response. In particular, the following prediction tasks attempt to accommodate the abilities required of a dialogue system to: 1) decide for an emotion triggering action, and 2) predict an emotional response to a trigger.

5·1 *Experimental Set Up*

Figure 7 illustrates the overview of the analyses and experiments on social-affective events. In the *emotional trig-*

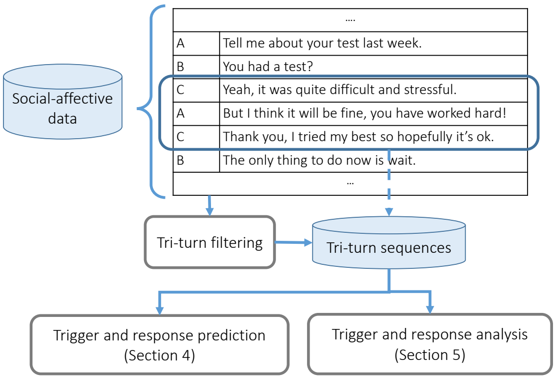

Social-affective data
....
A
Tell me about your test last week.
B
You had a test?
C
Yeah, it was quite difficult and stressful.
A
But I think it will be fine, you have worked hard!
C
Thank you, I tried my best so hopefully it's ok.
B
The only thing to do now is wait.
...
Tri-turn filtering
Tri-turn sequences
Trigger and response prediction (Section 4)
Trigger and response analysis (Section 5)

Fig. 7 Emotional triggers-responses prediction and analysis.

gers prediction, given the first and last turns of a triturn as an emotional response, we try to automatically predict what action takes place as the trigger, i.e., which dialogue act triggered this emotional response? On the other hand, in *emotional response prediction*, given the first two consecutive turns in a triturn, we try to predict the emotional response that will occur next, i.e., will the observed emotion dimension in the first turn rise, drop, or stay constant?

We partition the triturns from the constructed corpus with an 80:20 ratio to serve as training and test sets. On each triturn, we stack the features of the two respective turns to gather information of the context for predicting the event. That is, features from the first two turns are used to predict emotional response, and similarly for the first and last turns to predict emotional triggers.

The classification features comprise the acoustic information and the dialogue act of the respective two turns. These informations capture the emotion as well as the dialogue-related information of the triturn. Following the result from Chapter 3, we gather emotion-rich information from the speech by extracting the IS09 emotion challenge acoustic features [Schuller 09a] using the openSMILE toolkit [Eyben 10]. Stacking the two respective turns doubles the feature size, which is undesirable given the small amount of available data. To balance the ratio of instances to features, we perform correlation-based feature selection [Hall 99] and linear discriminant analysis.

As the classifier, we train a neural network with one hidden layer using Theano and the PDNN toolkit [Bergstra 11]. For each task, the input layer size follows the feature size, and the output layer follows the number of classes. For both tasks, we train 128 nodes in the hidden layer. We choose a small network architecture to compensate for the limited amount of data. The learning rate is 0.001 for the trigger prediction task, and 0.0003 for the response prediction tasks.

The class distributions on the test sets are as follows. For the dialogue act prediction, the test set contains 32 `stat`, 14 `back`, 5 `deny`, 4 `ack`, 2 `Qyno`, and 1 each of `opi`, `Qopn`, `Qwh`, `Qbck`, `thnk`,and `abdn`. There are 30 rise, 18 drop, and 15 constant samples for the arousal prediction task; and 29, 16, and 18 for valence. We determine the baseline chance level as described before in Section 3·1.

Table 7 Precision, recall, and F1 scores of social-affective events prediction tasks on test set in percent (%). The best F1 score on each task is boldfaced.

Task	Model	Prec.	Rec.	F1
Emotional triggers	Baseline	25.7	50.7	34.2
	Proposed	45.4	55.5	**47.5**
Arousal response	Baseline	22.6	47.6	30.7
	Proposed	53.9	52.3	**45.3**
Valence response	Baseline	21.1	46.0	29.0
	Proposed	58.1	58.7	**57.1**

5·2 *Experimental Result*

We summarize the evaluation of the triggers and responses prediction tasks in Table 7.

Emotional Trigger Prediction. The trained model is able to outperform the baseline chance level with an absolute improvement of 13.3%. However, further improvement is still necessary for a reliable prediction in real interaction. As elaborated on the analysis in Section 4·1, triggers of a certain emotion event in the data are not strongly characterized by the dialogue acts. In other words, the prediction accuracy is likely to be contributed by the other classification features, such as acoustic features and dialogue act of the context triturn. We found that the model is better at classifying instances that belong to the larger class. Collection of more data to increase the number of instances in all classes is necessary to improve the prediction performance.

Emotional Response Prediction. Both the valence and arousal predictions surpass the baseline chance level with considerable margins. The numbers show that prediction of valence events have the highest performance. This suggests a stronger pattern of action in discourse and speech characteristics when observing the change of valence, compared to that of arousal and the respective emotional triggers.

Inherently, there are numerous factors that leads to a change of emotion in a conversation. To properly recognize patterns for such events, a number of additional features are likely to be required to include other communication contexts. However, as the number of features rise, additional data would be needed to improve the model accuracy. As such, exploration into robust and efficient methods would also be necessary for future improvements.

6. Conclusions

In this paper, we presented a novel study on social-affective events in spontaneous human conversation collected from television talk shows. The proposed study differs from existing works in that it examines the change of emotion and its cause, on top of the occurrence of emotion itself. We carried out the emotion recognition task as the first step in studying social-affective aspects in human communication. Subsequently, we analyzed the constructed corpus in terms of social-affective events to uncover correlation between actions taken in discourse and the emotional response they trigger. Along with the social-affective events analysis, the result from the emotion recognition experiment provided information on emotion-rich speech features that is useful for proceeding to the emotional triggers and responses task. The experiments on predicting social-affective events resulted in models that surpass the chance rate level for both emotional triggers and responses prediction.

This new perspective of emotional events offers an approach in providing conversational agents and dialogue systems with social-affective capabilities: 1) to be able to decide for an emotion triggering action, and 2) to be able to predict an emotional response to a trigger. These two abilities are paramount to support emotionally intelligent agents in their interactions with a user. Moreover, this will endow agents with emotion appraisal knowledge to support its communication competences. Further improvement will be important for incorporation into dialogue systems, however we believe that this first step of exploration into new aspects of emotional occurrence provides a starting point for future developments and applications.

In future stages of the study, we hope to include more modalities of interaction in observing the dynamics of emotion, such as textual and visual features. A more detailed picture of the occurring events is highly potential in increasing the performance of the prediction models. As more features are considered in the study, it becomes important to balance the amount of data to avoid sparsity in the observation. Consequently, exploration into a more robust and efficient methods will be important for future improvements.

Acknowledgments

This research and development work was supported by the MIC/SCOPE #152307004 and JSPS KAKENHI Grant Number JP17K00237.

♢ References ♢

[Anagnostopoulos 15] Anagnostopoulos, C.-N., Iliou, T., and Giannoukos, I.: Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, *Artificial Intelligence Review*, Vol. 43, No. 2, pp. 155–177 (2015)

[Bergstra 11] Bergstra, J., Bastien, F., Breuleux, O., Lamblin, P., Pascanu, R., Delalleau, O., Desjardins, G., Warde-Farley, D., Goodfellow, I., Bergeron, A., et al.: Theano: Deep learning on GPUs with python, in *NIPS 2011, BigLearning Workshop, Granada, Spain* (2011)

[Chang 11] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 2, No. 3, p. 27 (2011)

[Douglas-Cowie 07] Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., et al.: The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data, in *Affective Computing and Intelligent Interaction*, pp. 488–500, Springer (2007)

[Egges 04] Egges, A., Kshirsagar, S., and Magnenat-Thalmann, N.: Generic personality and emotion simulation for conversational agents, *Computer Animation and Virtual Worlds*, Vol. 15, No. 1, pp. 1–13 (2004)

[El Ayadi 11] El Ayadi, M., Kamel, M. S., and Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition*, Vol. 44, No. 3, pp. 572–587 (2011)

[Ellsworth 03] Ellsworth, P. C. and Scherer, K. R.: Appraisal processes in emotion, *Handbook of Affective Sciences*, Vol. 572, p. V595 (2003)

[Eyben 10] Eyben, F., Wöllmer, M., and Schuller, B.: OPENsmile: the munich versatile and fast open-source audio feature extractor, in *Proceedings of the International Conference on Multimedia*, pp. 1459–1462ACM (2010)

[Eyben 16] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, *IEEE Transactions on Affective Computing*, Vol. 7, No. 2, pp. 190–202 (2016)

[Fleiss 71] Fleiss, J. L.: Measuring nominal scale agreement among many raters, *Psychological Bulletin*, Vol. 76, No. 5, p. 378 (1971)

[Hall 99] Hall, M. A.: *Correlation-based feature selection for machine learning*, PhD thesis, The University of Waikato (1999)

[Han 15] Han, S., Kim, Y., and Lee, G. G.: Micro-counseling dialog system based on semantic content, in *Natural Language Dialog Systems and Intelligent Assistants*, pp. 63–72, Springer (2015)

[Hasegawa 13] Hasegawa, T., Kaji, N., Yoshinaga, N., and Toyoda, M.: Predicting and Eliciting Addressee's Emotion in Online Dialogue, in *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*, pp. 964–972 (2013)

[Higashinaka 08] Higashinaka, R., Dohsaka, K., and Isozaki, H.: Effects of self-disclosure and empathy in human-computer dialogue, in *Proceedings of Spoken Language Technology Workshop*, pp. 109–112IEEE (2008)

[Hsu 03] Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al.: A practical guide to support vector classification (2003)

[Lasguido 14] Lasguido, N., Sakti, S., Neubig, G., Tomoki, T., and Nakamura, S.: Utilizing Human-to-Human Conversation Examples for a Multi Domain Chat-Oriented Dialog System, *IEICE TRANSACTIONS on Information and Systems*, Vol. 97, No. 6, pp. 1497–1505 (2014)

[Reeves 96] Reeves, B. and Nass, C.: *How people treat computers, television, and new media like real people and places*, CSLI Publications and Cambridge university press (1996)

[Ringeval 13] Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions, in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pp. 1–8IEEE (2013)

[Russell 80] Russell, J. A.: A circumplex model of affect, *Journal of* *personality and social psychology*, Vol. 39, No. 6, p. 1161 (1980)

[Scherer 01] Scherer, K. R., Schorr, A., and Johnstone, T.: *Appraisal processes in emotion: Theory, methods, research*, Oxford University Press (2001)

[Schuller 03] Schuller, B., Rigoll, G., and Lang, M.: Hidden Markov model-based speech emotion recognition, in *Proceedings of International Conference on Multimedia and Expo, 2003. ICME'03. Proceedings*, Vol. 1, pp. I–401IEEE (2003)

[Schuller 09a] Schuller, B., Steidl, S., and Batliner, A.: The INTERSPEECH 2009 emotion challenge, in *Proceedings of INTERSPEECH*, Vol. 2009, pp. 312–315 (2009)

[Schuller 09b] Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A.: Acoustic emotion recognition: A benchmark comparison of performances, in *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding, 2009*, pp. 552–557IEEE (2009)

[Skowron 13] Skowron, M., Theunis, M., Rank, S., and Kappas, A.: Affect and Social Processes in Online Communication–Experiments with an Affective Dialog System, *Transactions on Affective Computing*, Vol. 4, No. 3, pp. 267–279 (2013)

[Stolcke 00] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech, *Computational Linguistics*, Vol. 26, No. 3, pp. 339–373 (2000)

[Tielman 14] Tielman, M., Neerincx, M., Meyer, J.-J., and Looije, R.: Adaptive emotional expression in robot-child interaction, in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 407–414ACM (2014)

[Trigeorgis 16] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S.: Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204IEEE (2016)

[Zeng 09] Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 1, pp. 39–58 (2009)

[Zhao 16] Zhao, M., Adib, F., and Katabi, D.: Emotion recognition using wireless signals, in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pp. 95–108ACM (2016)

〔担当委員：黄 宏軒〕

Received May 29, 2017.

Author's Profile

Nurul Lubis

received the B.E degree (with distinction) in 2014 from Bandung Insistute of Technology, Indonesia and the M.Eng degree in 2017 from Nara Institute of Science and Technology (NAIST), Japan. She is currently a doctoral candidate at Augmented Human Communication Laboratory, NAIST, Japan. She is a recipient of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) scholarship. She was a research intern at Honda Research Institute Japan, Co. Ltd. Her research interest include affective computing, emotion in spoken language, and affective dialogue systems.

Sakriani Sakti

received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her Ph.D. degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR and U-STAR (2006-2011). In 2009-2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. From 2011, she has been an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015-2016, under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation". She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Koichiro Yoshino

received his B.A. degree in 2009 from Keio University, M.S. degree in informatics in 2011, and Ph.D. degree in informatics in 2014 from Kyoto University, respectively. From 2014 to 2015, he was a research fellow (PD) of Japan Society for Promotion of Science. From 2015 to 2016, he was a research assistant professor of the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST). Currently, he is an assistant professor of NAIST. He is also a researcher of PRESTO, JST, concurrently. He is working on areas of spoken and natural language processing, especially on spoken dialogue systems. Dr. Koichiro Yoshino received the JSAI SIG-research award in 2013. He is a member of IEEE, ISCA, IPSJ, and ANLP.

Satoshi Nakamura (Member)

is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampoli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.